

PROCESS VARIATIONS IN CMOS MEMORY DEVICES:
ANALYSIS, MITIGATION AND APPLICATION

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Sarah Qianying Xu

May 2014

© 2014 Sarah Qianying Xu

PROCESS VARIATIONS IN CMOS MEMORY DEVICES:
ANALYSIS, MITIGATION AND APPLICATION

Sarah Qianying Xu, Ph.D.

Cornell University 2014

Driven by the improvements on performance and cost, new generations of complementary metal oxide semiconductor (CMOS) memory devices such as SRAM, DRAM, and Flash have been aggressively scaled down to the deca-nanometer regime and beyond. Continued advancement of the CMOS technologies reduces the feature size and pitch and lowers the supply voltage to constrain power consumption. Cells and systems based on these devices are becoming more susceptible to process variations and transistor mismatches, causing various scaling challenges. On the other hand, researchers have recently demonstrated that inherent manufacturing variations can be exploited to authenticate an IC instance or generate unique secrets for each chip. This primitive is named physical unclonable functions (PUFs). In this work, we first study the impact of process variations on the 22nm prototype SRAM performance and stability caused by random dopant fluctuation (RDF), which is one of the dominant variation sources for sub-100nm devices. Hybrid SRAM-DRAM with cross capacitors is then designed and investigated for multi-bit storage, mismatch tolerance, and disturb stabilization capabilities, which help mitigate the severe scaling challenges in density, performance and stability. Finally, variation sources behind Flash PUF (FPUF) are decomposed and characterized to provide theoretical foundations for better implementation and utilization in security applications. Algorithms to improve the FPUF consistency and entropy are also discussed.

BIOGRAPHICAL SKETCH

In 1985, Sarah Qianying Xu was born into a family of physicists and engineers in Leshan, Sichuan, China. Her playgrounds as a young child were a bit of unusual, which included computer-filled control rooms and labs of gigantic apparatus. At that young age, despite of the early exposure to the scientific world, Sarah was more excited about becoming a business woman rather than a scientist due to all the television she had been watching.

However, it turns out the apple doesn't fall far from the tree after all. Always a little boyish and quirky at heart, young Sarah's hobbies revolved with reading science fictions, building model toys, and earning some extra allowances from passing strangers by exhibiting her premature violin skills outside the front door. One day in school, a city-wide competition was announced where students could enter their home made crystal radios for prizes. Although having no idea what a crystal radio was at that time, Sarah saw the opportunity to gain her long-dreamed roller blades. After some careful research and a lot of question asking, the 12-year old bought her very first bag of diodes, learned big words like piezoelectric, and put her crafting skill to work with a soldering gun. The moment she heard the music coming out of the speakers on her partially burned breadboard, the joy of experiencing something magical rushed through her heart. That radio did not only win her second place in the competition, but also later on intrigued her to pursue a B.S. degree of Applied Physics with minoring in Electrical Engineering from Queen's University in Kingston, Ontario, Canada.

As first-ranked in her class and received multiple awards every year, although only a junior undergraduate student, Sarah was selected to become a research intern to join the *project in Canada to search for super symmetric objects* (PICASSO) supervised by Professor Anthony Noble. The experiments were held in SNOLAB,

which is one of the world's deepest underground lab facilities, and the place where she first learned the protocols of conducting research in a cleanroom. She received her bachelor degree with First Class Honor in 2008, and decided to continue her academic pursuits by applying for a doctorate study in the field of electrical engineering.

After all these years, the joyful and amazed feeling from that little radio project never escaped her heart. Hoping to gain more hands-on research opportunities and in-depth understanding of various electronic devices, Sarah joined research group of Prof. Edwin C. Kan in the School of Electrical and Computer Engineering at Cornell University in 2008. Her research has centered on combined simulation and experimental studies of CMOS memory devices, including SRAM, DRAM and Flash memories, in order to gain better understandings of the process variation effects on these aggressively scaled devices. She hopes her works on hybrid SRAM-DRAM circuits and characterizations of Flash Physical Unclonable Functions will show new directions on mitigating and utilizing the inherent variation issues for these ubiquitous memory devices in today's embedded systems and modern integrated circuits.

To my dearest mother, Li Liu

ACKNOWLEDGMENTS

Graduate school is an experience unlike any other, at least for me. Although it is a place known for developing intellectually and attaining valuable knowledge, I feel the most drastic impact of the past five and half years at Cornell is how I grew and matured into a professional person, not only academically, but also in everyday life. I believe I learned about the importance of curiosity, perseverance, and integrity during my research process, developed my communication skills as a teaching assistant, and trained my meticulousness through the time spent inside Cornell Nanofabrication Facility (CNF). However, just like what Helen Keller once said: “Alone we can do so little; together we can do so much”, without the guidance and support from all the people I had the pleasure of knowing during my study, I would never be able to stand where I am today. This is why I would like to thank everyone who accompanied me through this bitter sweet journey.

My dear advisor, Prof. Edwin C. Kan, is the first person I’d like to express my sincere appreciation to. I applied Cornell in hopes of studying specifically in his group, and I’m more than honored that he accepted me as one of his students. He’s known for his brilliance as a researcher, dedication as a teacher, and most importantly, his extraordinary accommodation and understanding of personal difficulties. As a graduate student, besides the sleepless nights and the head scratching moments, life can throw many other unexpected things at you, and Prof. Kan is always willing to go extra miles to help people out. I cannot remember how many times I went to his office, sad and frustrated, like when my grandfather passed away or when I didn’t know if accepting a specific job offer was the smart decision, and chatting with him always calmed me down. I feel he is not only my academic advisor, but also someone I look up to as a lifelong role model.

I’m also very grateful for Prof. G. Edward Suh and Prof. R. Bruce van Dover,

who not only diligently served on my Ph.D committee, but also guided me through many projects with inspiring feedbacks. I would like to thank Prof. Sandip Tiwari and Prof. Farhan Rana, who taught me many things that are crucial to my research. I'd also like to extend my gratitude towards Prof. Aaron B. Wagner, who helped me out with a lot of last minute questions. Talking with them always provided refreshing new perspectives and solutions for fundamental problems, which kept my research on track.

Besides the professors, people I spent most of my time with are my office mates, who supported me every day with their enlightening conversations and relaxing chats. I would like to thank Jonathan Shaw for his mentoring from the beginning of my study, helping me learning all the necessary skills in the bleak CNF cleanrooms. I'd also like to show my gratitude towards Jaegoo Lee and Hassan Raza for patiently teaching and guiding me through my first year of study. I'm extremely thankful towards former and current students from Kan's group, who include Shantanu Rajwade, Fan Yu, Krishna Jayant, Xiaoyang Li, Keith Lyon, Nini Muñoz, Phillip Gordon, Joshua Phelps, and Yinglei Wang, for their company and delighted times spent together, especially Kshitij Auluck, Adrian Lieh-Ting Tung, and Yunfei Ma, who kept me joyful during the harsh Ithaca winters, and anchored me inside when the beautiful summer tried to lure me out. I'm very appreciative towards Wing-Kei Yu, who I had the pleasure of collaborating so many times. I also treasure the friendship I had with fellow graduate students, Xiao Wang, Xuan Zhang, Bo Xiang, and Haining Wang, for their help and support.

I would like to acknowledge the staffs from CNF, Nanobiotechnology Center, and Cornell Center for Materials Research, particularly Aaron Windsor, Rob Ilic, Phil Infante, Steve Kriske, and Jonathan Shu for their patience during my training and assiduous maintenance of all the equipment I have used. Special thanks to my internship manager Phil Oldiges, who gave me a very smooth introduction to the

industry world. Finally to my lovely mother Li Liu, who continues to give me all her unconditional love and support; without her, it would be extremely difficult for me to be mentally sane after all the bugged programs and frustrating experiments.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgments	vi
Table of Contents	ix
List of Figures	xiii
List of Tables	xviii
List of Abbreviations	xix
List of Symbols	xxii
 CHAPTER 1: INTRODUCTION	 1
1.1 <i>CMOS Device Variations and Scaling Effects</i>	1
1.1.1 Background on CMOS Variability.....	1
1.1.2 Manufacture Variations: Spatial and Layout.....	2
1.1.3 Manufacture Variations: Intrinsic Device Variations.....	4
1.1.4 Variation in the Field: Stress-induced Variations.....	6
1.1.5 Dynamic Variations: Random Telegraph Noise.....	9
1.2 <i>Major Variation Sources for Sub-100nm Devices</i>	10
1.3 <i>Impact of Process Variations on Memory Devices</i>	13
1.3.1 Circuit and System Performance Degradation in SRAM.....	13
1.3.2 Increased Threshold Voltage Variations for Flash Memory....	13
1.3.3 Mitigate the Impact of Process Variations.....	14
1.4 <i>Chapter Organization</i>	15
<i>REFERENCE</i>	16
 CHAPTER 2: IMPACT OF RDF ON 22NM SRAM NOISE MARGINS.....	 24

2.1	<i>Abstract</i>	24
2.2	<i>Introduction</i>	24
2.3	<i>Individual Transistor Simulation</i>	25
2.3.1	Simulation and Analysis Methodologies.....	25
2.3.2	Threshold Voltage Distributions.....	26
2.3.3	Transport Characteristics.....	29
2.4	<i>Mixed Mode SRAM Simulations</i>	30
2.4.1	Mixed Mode Simulation Methodologies.....	30
2.4.2	SNM and SINM.....	31
2.4.3	Tail Distributions of SNM and SINM.....	34
2.5	<i>Conclusion</i>	34
	<i>REFERENCE</i>	36

CHAPTER 3: HYBRID SRAM-DRAM WITH CROSS CAPACITORS FOR MULTI-BIT STORAGE AND DISTURB STABILIZATION.....38

3.1	<i>Abstract</i>	38
3.2	<i>Introduction and Motivation</i>	38
3.3	<i>Hybrid Cell Structure</i>	39
3.4	<i>SRAM-DRAM Hybrid Cell Operations</i>	40
3.4.1	Regular Operations.....	40
3.4.2	Context Switching between SRAM and DRAM.....	41
3.4.3	Retention.....	43
3.5	<i>Prototype Measurements</i>	44
3.6	<i>Disturb Stabilization</i>	46
3.6.1	Disturb Stabilization Analysis.....	46
3.6.2	Multi-bit Storage and Disturb Stabilization.....	51

3.7	<i>Transient Noise Margin</i>	51
3.8	<i>Conclusion</i>	55
	<i>REFERENCE</i>	57

CHAPTER 4: IMPLICATIONS OF VARIATION SOURCES ON FLASH PHYSICAL UNCLONABLE FUNCTION DESIGN AND USAGE.....59

4.1	<i>Abstract</i>	59
4.2	<i>Introduction</i>	59
4.3	<i>Overview on Flash Physical Unclonable Function (FPUF)</i>	61
4.4	<i>Manufacturing Variations in Flash PUF</i>	63
4.4.1	Layout Variations.....	63
4.4.2	Spatial Variations.....	65
4.4.3	Intrinsic Variations.....	67
4.4.4	Implications in FPUF design.....	70
4.5	<i>Variations in the Field</i>	72
4.5.1	Stress-induced Variation.....	72
4.5.2	Random Telegraph Noise.....	73
4.5.3	Block-level Erase Effect.....	75
4.5.4	Implications for FPUF design	76
4.6	<i>Quantization Algorithms and Respective Maximum Entropies</i>	78
4.6.1	Short-and-Long Bits.....	78
4.6.2	Pair-wise Comparison	79
4.6.3	Cell-wise and Pair-wise Entropies.....	80
4.7	<i>Future Possibility for Data Encryption</i>	81
4.8	<i>Conclusion</i>	83
	<i>REFERENCE</i>	84

CHAPTER 5: PROBING THE ORBITAL LEVELS OF ENGINEERED FULLERENIC MOLECULES FROM A NONVOLATILE MEMORY CELL	88
5.1 <i>Abstract</i>	88
5.2 <i>Introduction</i>	88
5.3 <i>Experiments</i>	89
5.4 <i>Discussion</i>	90
5.5 <i>Conclusion</i>	95
<i>REFERENCE</i>	96
CHAPTER 6: CONCLUSION	98
6.1 <i>Summary of Major Contributions</i>	98
6.2 <i>Suggestions for Future Work</i>	99
6.2.1 <i>Flash Rank Modulation</i>	99
6.2.2 <i>Trial Implementation with Partial Program Operation</i>	101
<i>REFERENCE</i>	104

LIST OF FIGURES

Figure 1.1	Classification of various sources for CMOS devices variability	2
Figure 1.2	Schematic of Random dopant fluctuation (RDF) and line edge roughness (LER) in a MOS transistor	5
Figure 1.3	Schematic of stress-induced leakage current (SILC) mechanism and illustration on trap assisted tunneling mechanism	7
Figure 1.4	Schematic of random telegraph noise mechanism, measured signal and its power spectrum signature with $1/f^2$ characteristic	9
Figure 1.5	RDF and LER contributions to σI_{off} for sub-100nm regime	11
Figure 1.6	RDF induced σV_{th} scaling trend for sub-100nm MOS devices	11
Figure 2.1	Cross section of the prototype 22nm NFET device with an instance of Monte Carlo introduced discrete random dopant fluctuation in the channel	26
Figure 2.2	Linear threshold voltage distribution and saturation threshold voltage distribution of RDF Pull-down NFETs referenced to nominal device results	26
Figure 2.3	Saturation ΔV_{th} vs linear ΔV_{th} referenced to the nominal device values for all three devices: RDF, Continuum and WF	28
Figure 2.4	DIBL vs linear V_{th} for all three cases. DIBL of WF devices only showed 1mV fluctuations which was caused by the granularity of data extraction	28
Figure 2.5	Linear overdrive current vs linear ΔV_{th} . WF devices showed no V_{th} dependency due to unvaried channel doping profile	29
Figure 2.6	Potential barrier at threshold voltage bias for nominal and RDF devices	

	having the same number of channel dopant	30
Figure 2.7	SRAMs composed of WF and Continuum devices show similar SNM and SINM, and both are larger than RDF cells	33
Figure 2.8	SNM and SINM for SRAMs composed of RDF, WF and Shifted WF devices	33
Figure 2.9	Lower tail distributions for SNM and SINM of 22nm prototype SRAM with both normal and Poisson curve fittings	34
Figure 3.1	Schematics of the SRAM-DRAM hybrid cell	40
Figure 3.2	Critical EN rising time with various DRAM capacitances	41
Figure 3.3	SNM and SINM for SRAMs composed of RDF, WF and Shifted WF devices	42
Figure 3.4	RESTORE operation testing sequence and oscilloscope measurements of BL, BLb, WL and the external enable signal	43
Figure 3.5	STORE operation testing sequence and oscilloscope measurements of BL, BLb, WL and the external enable signal before double buffer sharpening	44
Figure 3.6	BL voltages at various retention times	45
Figure 3.7	Range of enable rising time for achieving a successful RESTORE operation	46
Figure 3.8	EN rising time range for the hybrid cell with 5fF cross capacitance that can achieve a successful STORE operation	47
Figure 3.9	Read current noise margins (SINM) at 200mV inverter mismatches for both regular SRAM and the SRAM-DRAM hybrid cell with 5fF cross capacitor	48
Figure 3.10	Conventional read current noise margin measurements for the hybrid cell	49

Figure 3.11	The N-curve extracted through BLb, its corresponding SRAM states and leakage current through cross capacitor with $E_N=0V$	50
Figure 3.12	The N-curve extracted through BLb, its corresponding SRAM states and leakage current through cross capacitor with $E_N=1.5V$	50
Figure 3.13	A sample illustration of the linear current source at a ramping speed of $10\mu A/ns$	51
Figure 3.14	Transient read current noise margin (TINM) under various inverter mismatches and noise injection speeds with no cross capacitor	52
Figure 3.15	TINM differences with and without 5fF cross capacitor under various inverter mismatches and noise injection speeds	53
Figure 3.16	TINM for various cross capacitances with 200 mV inverter pair mismatch and 10ns noise duration	54
Figure 3.17	SINM and TINM with various inverter mismatches and capacitances for 10ns noise duration	55
Figure 4.1	Flash memory cell schematics and equivalent operating circuit	62
Figure 4.2	Systematic layout variations in two Flash chips with the same part number	64
Figure 4.3	Fluctuations across 4000 blocks in three chips of the same part number but from different production lots and wafers	66
Figure 4.4	Correlations of Flash fingerprint measurements obtained from the same page and with different pages	66
Figure 4.5	Example of FPUF distribution and translated threshold voltage distribution with Normal fitting	68
Figure 4.6	Translated threshold voltage distributions from three technology generations and the respective normal fittings	69
Figure 4.7	Expected scaling trend of RDF induced threshold voltage variations	

	and experimental translated ones from Flash chips measurements vs different technology nodes	70
Figure 4.8	P/E stress effect on average partial program numbers and FPUF correlation coefficients	72
Figure 4.9	Bit-wise fluctuations for multiple FPUF measurements	73
Figure 4.10	Power spectral density of bit-wise fluctuations and corresponding $1/f^x$ coefficients.....	74
Figure 4.11	Bit-wise fluctuation correlation distributions obtained with regular erase or fixed erase	75
Figure 4.12	Percentage of partial program number variations of FPUF responses obtained between single measurements and between averaged measurements	77
Figure 4.13	Average percentage bit-wise fluctuations between FPUFs with full erase and fixed erase against number of averaged measurements	77
Figure 4.14	Simple short-long quantization algorithm for FPUF	79
Figure 4.15	Probabilities of pair-wise comparison yield value “1” across all pages in a block	80
Figure 4.16	PMF for single Flash cell pair-wise partial program number distribution.....	81
Figure 5.1	Design splits of the EFM gate stack	89
Figure 5.2	Room-temperature and low-temperature (10K) high frequency CV measurements on toluene-spin-coated C ₆₀ -PCBM and C ₇₀ -PCBM in comparison with the control sample S1.....	91
Figure 5.3	Band diagram and schematic of the MOS capacitor gate stack...	92
Figure 5.4	ΔE_{EFM} extraction from ΔV_{FB} for C ₆₀ -PCBM and the band diagram from calculated C ₆₀ -PCBM molecular orbital	93

Figure 6.1	Conventional Flash memory programmed states for SLC and MLC devices	99
Figure 6.2	SLC rank modulation algorithm utilizing partial program operations	101
Figure 6.3	Read results for rank vector (5, 6, 3, 7, 4, 1, 8, 2) with 200 samples	103

LIST OF TABLES

Table 2.1	Standard deviations of threshold voltages for RDF devices at both linear ($\sigma V_{t_{lin}}$) and saturation ($\sigma V_{t_{sat}}$) regions using a sample size of 200... ..	27
Table 2.2	SNM and SINM variations for SRAMs composed of RDF, Continuum, WF and shifted WF devices	32
Table 4.1	Bit capacity per area comparisons for various PUFs	63
Table 4.2	Flash chips tested for FPUF investigations	63
Table 4.3	Diehard tests on FPUFs with and without systematic components	71
Table 4.4	Simple example of utilizing quantized FPUF for data encryption	82
Table 6.1	Relationship between group size N and resulted equivalent bit capacity	100
Table 6.2	Sample program and read operation results for rank modulation	102

LIST OF ABBREVIATIONS

1T-1C	One Transistor and One Capacitor
2T-1C	Two Transistors and One Capacitor
6T	Six-Transistor
ALD	Atomic Layer Deposition
BL	Bitline
BTI	Bias Temperature Instability
CD	Critical Dimensions
CMOS	Complementary Metal Oxide Semiconductor
CNL	Charge Neutrality Level
CV	Capacitance-Voltage
DIBL	Drain-Induced Barrier Lowering
DRAM	Dynamic Random Access Memory
EFM	Engineered Fullerenic Molecule
EN	Enable
FBB	Forward Body Bias
FET	Field Effective Transistor
FPUF	Flash Physical Unclonable Function
GB	Grain Boundary
HCI	Hot Carrier Injection
HOMO	Highest Occupied Molecular Orbital
IC	Integrated Circuit
L2L	Lot to Lot
LER	Line Edge Roughness
LUMO	Lowest Unoccupied Molecular Orbital

LWR	Line Width Roughness
MOS	Metal Oxide Semiconductor
NBTI	Negative BTI
NVM	Non-Volatile Memory
NWE	Narrow-Width Effect
OPC	Optical-Proximity Correction
OPE	Optical Proximity Effect
PBTI	Positive BTI
PD	Pull-Down
PG	Pass-Gate
PSG	Poly-Silicon Granularity
PSM	Phase-Shift Masking
PU	Pull-Up
PUF	Physical Unclonable Function
RDF	Random Dopant Fluctuations
RTN	Random Telegraph Noise
RTS	Random Telegraph Signal
SCE	Short-Channel Effect
SILC	Stress-Induced Leakage Current
SINM	Static Read Current Noise Margin
SNM	Static Noise Margin
SRAM	Static Random Access Memory
TAT	Trap Assisted Tunneling
TCAD	Technology Computer Aided Design
TDDDB	Time-Dependent Dielectric Breakdown
TINM	Transient Current Noise Margin

W2W	Wafer to Wafer
WF	Work Function
WID	Within a Die
WL	Word Line

LIST OF SYMBOLS

α_{CP}	Flash gate coupling ratio
ϵ_{conl}	Control oxide permittivity
ϵ_{ox}	Oxide permittivity
σI_{off}	Standard deviation of off current
σN	Standard deviation of dopant distribution
σV_{th}	Standard deviation of threshold voltage
τ_c	Capture time constant
τ_e	Emission time constant
ΔV_{FB}	Flatband voltage shift
C_{l_3D}	3D substrate-NC coupling capacitance in the unit cell
C_{2_3D}	3D NC-gate coupling capacitance in the unit cell
C_{FG}	Self-capacitance of the charge storage node
C_{NC}	NC unit cell diameter
E_{CH}	Single-electron charging energy
E_{C60_HOMO}	C ₆₀ HOMO energy
E_{C60_LUMO}	C ₆₀ LUMO energy
E_{Si_C}	Silicon conduction band edge energy
I_D	Drain current
Iodline	Linear overdrive current
L_{eff}	Effective channel length
N	Dopant number/concentration
Q	Number of electrons
R_{3D}	3D channel-control factor
SoS	Standard deviation of standard deviation

T_{ox}	Oxide thickness
V_{th}	Threshold voltage
V_G	Gate voltage
W_d	Depletion layer width
W_{eff}	Effective channel width

CHAPTER 1

INTRODUCTION

1.1 CMOS Device Variations and Scaling Effects

1.1.1 Background on CMOS Variability

Several decades of rapid growth in electronic industry has been predominately driven by the drastic device scaling in complementary metal oxide semiconductor (CMOS) technology. Because of its limiting effects in both performance and overall chip yield, variability in CMOS device has always been a key topic in circuit and system designs [1]. New generations of CMOS memory devices, such as static random access memory (SRAM), dynamic random access memory (DRAM), and Flash non-volatile memory (NVM), have been aggressively scaled down to the deca-nanometer regime and beyond, due to requirements on revenue and performance [1-3]. This decrease in device feature size results in ever increasing processing complexities, thus introducing new variation sources that can no longer be overlooked.

Variability in CMOS memory device refers to the amount of deviation from the intended value of a particular design parameter. In the past, the variability mainly came from imperfect control of the fabrication process, and the majority of the device performance variations exist between lot to lot (L2L) and wafer to wafer (W2W) [4]. Because device feature sizes are gigantic at that time (over microns), variation within a die (WID) was relatively small. Sources of variation were more of a global nature rather than local fluctuations. However, radical scaling of the CMOS devices has changed the statistical nature of the variation sources significantly. Non-lithographic issues such as WID layout dependent variations and device to device fluctuations have become the important contributors to the overall device reliability problems [4]. For instance, intrinsic device variations, such as discrete random dopant fluctuation (RDF)

and line edge roughness (LER), are dominating the process variation for sub-100nm devices. Thus, manufacturing problems, together with atomistic and quantum-mechanical limitations at nanometer regime [5], introduce device variability that can no longer be controlled through advancement in lithography [6] or mitigated through conventional corner-based design approaches [1]. 3-D “atomistic” studies are required in the TCAD simulations [5] in order to gain better understandings of the new major variation sources. Novel approaches in dealing with and utilizing CMOS device variability have become main topics for current researchers and designers [7-9].

This study will be focusing on understanding, mitigating and utilizing variability in CMOS memory devices. An overview of the CMOS variations is presented in this chapter, and a simple classification is illustrated in Fig. 1.1. From the graph, the variation sources include ones originated from the manufacture process, such as spatial, layout and random intrinsic device variations, as well as variations in the field, such as stress-induced variations and dynamic variations.

1.1.2 *Manufacture Variations: Spatial and Layout*

Manufacture variation refers to the device fluctuations introduced during the

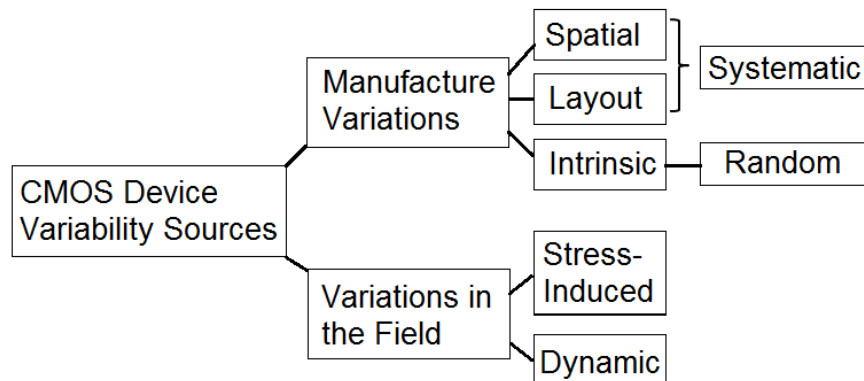


Figure 1.1 Classification of various sources for CMOS devices variability.

original fabrication process. For the purpose of circuit and system design, the manufacturing variations can be classified into two main categories: systematic and random. Systematic differences in CMOS devices usually display a consistent mean value shift in the sensitive design parameters originated through spatial and layout dependent variations [10], which can be caused by numerous non-uniformity that are common for the wafer manufacture process using the same recipe or in the same environment. These spatial and layout variations will result in identical groups of devices to behave differently at different die or wafer locations, and can often be controlled to an extent through better fabrication techniques and resolutions [10]. Random variations, on the other hand, come from intrinsic device parameters that cannot be fully eliminated through current design techniques, and can result in unpredictable variations between identical devices regardless of their locations. While these relationships are more of a common knowledge for the process people, they are usually not well explained to the circuit and system designers. In order to better account for the variability in the design process, it is necessary to distinguish systematic deviations in parameter values from ones originated from random sources [11].

Spatial variations can be generated through the deposition process, photo-resist spinning, reticle imperfections and etching [10, 12-14], and the effects often consist of similar parameter gradient across the wafer fields, such as gate dimensions and layer thickness [4,6]. Layout dependent variations, on the other hand, can cause groups of identical devices adjacent to each other with two different layouts to exhibit vastly different characteristics, even when the spatial factor should be negligible [10]. At the same time, instances with same layout often show high correlations in their electronic performances, regardless if they are close or not. Factors contributing to the layout variations include optical-proximity correction (OPC), phase-shift masking

(PSM) layout-induced strain and well-proximity effect [15], which can manifest as fluctuations in channel length, channel width, layer thickness, resistivity, average doping density and body effect [14, 16-18]. Therefore, spatial variation can be isolated by looking at the parameter gradient across wafer field between several wafers, and layout induced variability can be identified through consistent systematic parameter fluctuations between chips with the same layout designs. In general, spatial and layout variability are relatively predictable, and can thus be modeled as a function of deterministic factors such as surrounding topological environment and layout structures [19].

1.1.3 *Manufacture Variations: Intrinsic Device Variations*

Different from the systematic variations caused by spatial and layout aspects, local random variability often arises from the intrinsic device parameters that cannot be fully duplicated through current fabrication techniques. Random variations can cause significant deviation between identical devices, even if their locations are close to each other. Major intrinsic variation sources include discrete random dopant fluctuations (RDF), line edge roughness (LER), poly-silicon granularity (PSG), and line width roughness (LWR) [15]. These individual device differences are unpredictable and usually quantified through statistical studies [20-24]. As a result, process corner-based design methodologies, where verification is only performed at a small number of tail devices, may become insufficient for correct performance considerations [1].

Fig. 1.2 illustrates the effect and location of RDF and LER in a metal oxide semiconductor (MOS) device. RDF refers to the random fluctuation of the relatively small number of dopants and their discrete microscopic arrangement in the channel region, which will lead to significant variations in threshold voltage (V_{th}) and drive current (I_D) [5 [27]. Random dopant fluctuation is particularly serious for minimum

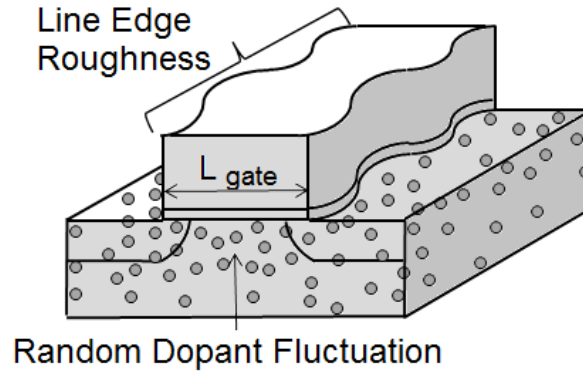


Figure 1.2 Schematic of Random dopant fluctuation (RDF) and line edge roughness (LER) in a MOS transistor.

geometry device and cells such as Flash and SRAM, because aggressive scaling of the CMOS devices can drastically reduce the number of dopant atoms in the channel region of these transistors.

LER is introduced through patterning with non-ideal gate edges [25-27], as shown in the Fig 1.2. When photolithography uses light source with wavelength much larger than the minimum feature size [21], gate variation due to LER can be exacerbated. Diffraction of the light can cause additional distortions due to optical proximity effect (OPE) [27-29], and is one of the major concerns when defining critical dimensions (CD) [27, 30, 31]. As scaling continues into the deca-nanometer regime, LER will not scale accordingly, and become an increasingly larger fraction of the overall process variations.

For devices that incorporate poly-silicon gate, PSG increases as technology scales, and the non-uniformity in gate doping exacerbates as well [7, 32, 33]. This is because gate dopant diffusion can be enhanced along the grain boundaries (GBs), leading to localized penetration of dopants through the gate oxide, migrating into the channel region. The most significant effect of PSG is thought to be Fermi-level

pinning between grains due to defect states, and thus introduces substantial gate local potential variations [15].

LWR or channel width variation is similar to LER but focuses on the transistor channel width variability caused by lithography limitations. Devices with LWR usually suffer from narrow-width effects (NEW) [34]. However, since the width of the device is typically much larger than the gate length, the contribution of LWR on is often considered not as significant as LER [34].

There are also several other factors that can contribute to the overall device variability, although currently not important enough to compete with aforementioned variation sources. Some of them include mobility fluctuations and gate oxide thickness (T_{OX}) variations [15]. Mobility fluctuation can change a transistor's drive current, which is the result of several complex physical mechanisms, such as fluctuations in effective fields, fixed oxide charges, doping, and inversion layers [25]. Gate oxide fluctuation can affect many device characteristics, in particular V_{th} . However, since T_{OX} is one of the best controlled parameters in CMOS devices, its effect on V_{th} variation is not substantial.

1.1.4 Variation in the Field: Stress-induced Variations

In addition to the spatial, layout and intrinsic device fluctuations which are time-independent variation sources, highly scaled MOS based devices are subject to degradation over time due to long-term stress and device aging [1]. This kind of variability caused by field operation is referred to as variation in the field. Stress-induced variability factors such as stress-induced leakage current (SILC) and bias temperature instability (BTI) are generated through significant amount of defects inside the device structures, which can change their electrical characteristics and compromise the specifications required for the respective electronic components [35].

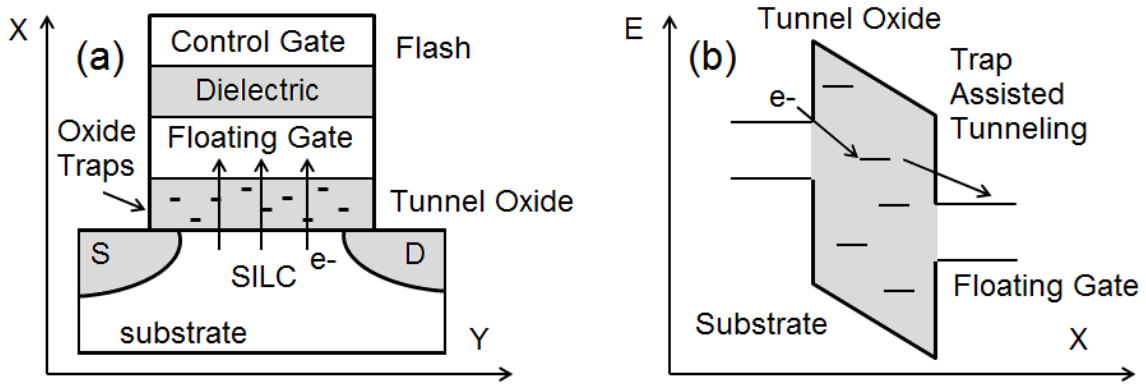


Figure 1.3 (a) Schematic of stress-induced leakage current (SILC) mechanism in Flash. (b) Illustration on trap assisted tunneling mechanism.

These degradations in integrity and reliability of MOSFETs and Non-volatile memory (NVM) devices have become a topic of great importance for modern day researchers and designers [35, 36].

SILC refers to the increase in gate leakage current due to the two-step trap assisted tunneling (TAT) mechanism enabled by the available energy states, as shown in Fig. 1.3. These energy states are caused by the stress-induced defects, which are generated when carriers are injected across the thin tunnel oxide during repetitive operations. High field stress can produce damages and break atomic bonds in the molecular structure of the silicon dioxide, thus generating available states in the gap of the oxide energy band [37]. These defect states are commonly known as the bulk oxide traps. TAT is thus enabled by the presence of these stress-induced defects, and gate current can be considerably enhanced in the MOS structure, particularly at low field, which is a major reliability concern for both transistors and Flash devices [37]. More specifically, stress induced oxide traps can generate conductive percolation paths that yield to the discharge of the FG, which can severely hamper the retention

characteristics of the floating gate memory devices as well [38, 39].

An additional degradation mechanism also exists in MOS devices, which is called bias temperature instability (BTI). While SILC has been studied extensively, BTI has received relatively less attention, even though it is one of the earliest identified reliability problems [40, 41]. This is not because BTI has been fully understood, but because in modern MOS integrated circuit (IC), this effect has been greatly reduced through process control [42]. However, recent experimental results [43] have shown that the BTI can still make a considerable contribution to the degradation of MOS devices with small feature size, which is why it is worth mentioning about.

Different from SILC, where defects are usually generated near the silicon and oxide interface [44, 45] and/or in the bulk gate oxide [46], BTI is induced within a few tens of nanometers below the interface [41], and the resulted V_{th} shift can recover slightly after the stressing condition is removed. However, it has been found that the recovering time is in the range of millisecond to second range, which complicates the process of accurately measure the BTI effects. Therefore, it is quite difficult to experimentally separate BTI from SILC. BTIs are usually categorized as negative BTI (NBTI) and positive BTI (PBTI). Although NBTI are more studied due to its severity in PMOS devices [47], effects of PBTI cannot be easily ignore either [42].

In addition to all these prominent time-dependent degradation factors, there are still other aging effects existed for MOS devices, such as hot carrier injection (HCI) [35] and time-dependent dielectric breakdown (TDDB) [48].

1.1.5 Dynamic variations: Random Telegraph Noise

Random telegraph noise (RTN), also known as random telegraph signals

(RTS) [49, 50], is an additional random intrinsic variation source that is universal to all CMOS devices. Significantly different from RDF, RTN is a time-dependent phenomenon with relatively large time constants and much fewer charges involved [49]. RTN is caused by the random trapping and de-trapping of channel carriers in the dielectric traps happening at the oxide interface, as shown in Fig. 1.4 (a) [50]. Fig. 1.4(b) illustrates the resulted ΔV_{th} manifested as fluctuations between two or more distinct levels, depending on the trap energy [35]. Taking care of the RTN noise have always been an important part for analog and radio frequency circuits [51]. As CMOS technology scales down, role of RTN increases significantly for minimum feature devices such as SRAM [49, 50, 52] and Flash [53] due to the reduction in the number of channel carriers.

RTN behavior can be identified through its unique $1/f^2$ spectrum density (PSD) characteristic especially at low frequencies [54, 55], as depicted in Fig. 1.4 (c) . Capture and emission time constants (τ_c and τ_e) can be used to acquire information of the trap energy levels [56, 57]. With multi-level RTN due to the activation of

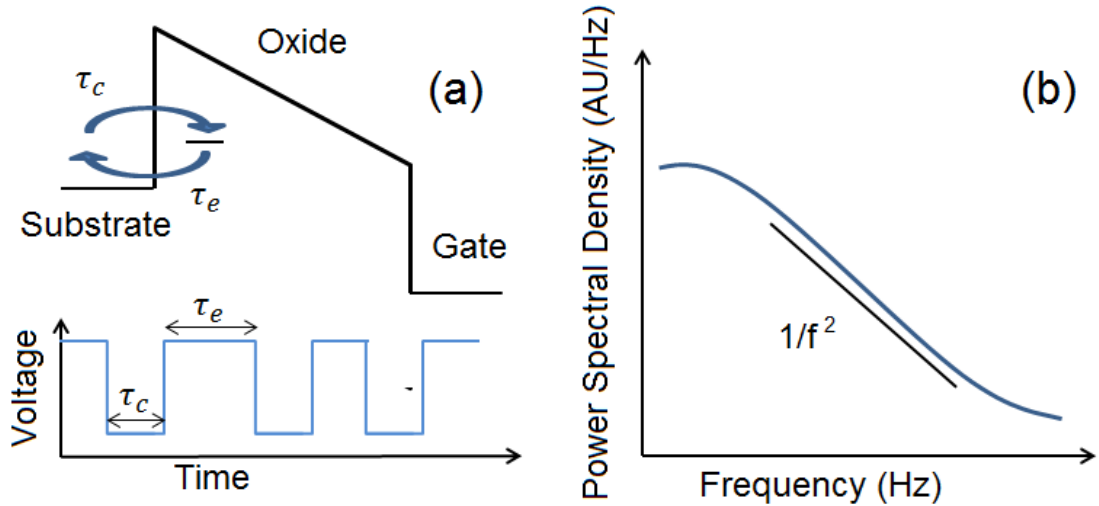


Figure 1.4 Schematic of (a) random telegraph noise mechanism, (b) measured signal and (c) its power spectrum signature with $1/f^2$ characteristic.

multiple traps near the quasi-Fermi level, the PSD usually display the $1/f$ signature with frequency power coefficient ranging between 1 to 2 [54, 58].

1.2 Major Variation Sources for Sub-100nm Devices

Random variation sources such as oxide layer thickness change, dopant fluctuations, and other lithographic dimension deviations have been examined extensively through literature [1, 4, 6, 10]. Among these factors, RDF and LER [59] are considered to be the predominant intrinsic variation sources in today's electronic devices, since they contribute to the majority of the threshold voltage fluctuations in sub-100nm CMOS devices. Many groups have attempted to estimate the total percentage variability contributions of these two factors by comparing measured data to simulation [5, 60]. It is reported that RDF alone can contribute to approximately 65% of the overall 65nm NMOS σV_{th} [64]. Circuits and systems that rely on minimum feature devices or relative device matching may have their functionality and performances severely affected by these kinds of process variations, such as SRAM and Flash memory [65].

As illustrated in Fig. 1.5, for sub-100nm technologies, RDF is the dominant factor between the two due to V_{th} sensitivity to the drastically reduced channel dopant number. Due to the discreteness of atoms, a statistical random fluctuation is associated with the number of dopant within a limited volume, which follows a Poisson distribution [25]. The severity of the RDF effect on V_{th} can thus be examined.

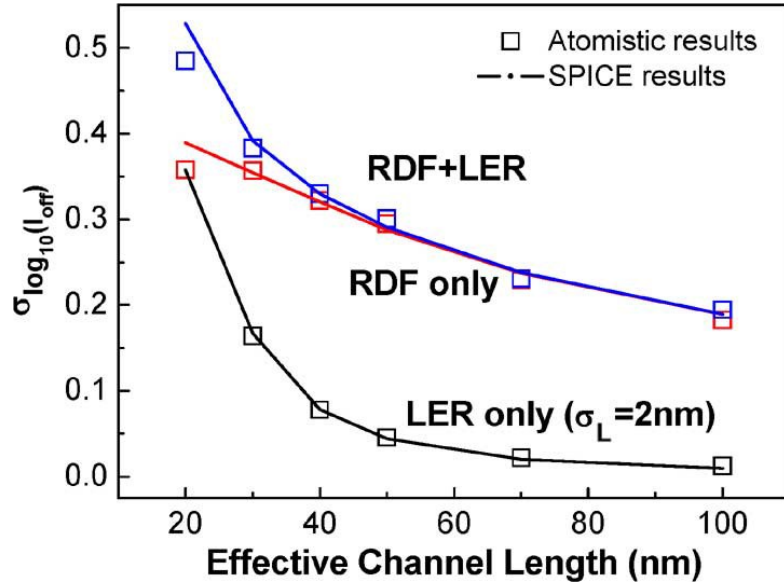


Figure 1.5 RDF and LER contributions to σI_{off} for sub-100nm regime, which is directly related to σV_{th} . RDF is dominating before device channel length reduces to 40nm; after 40nm, LER starts to become a serious contender [61-63].

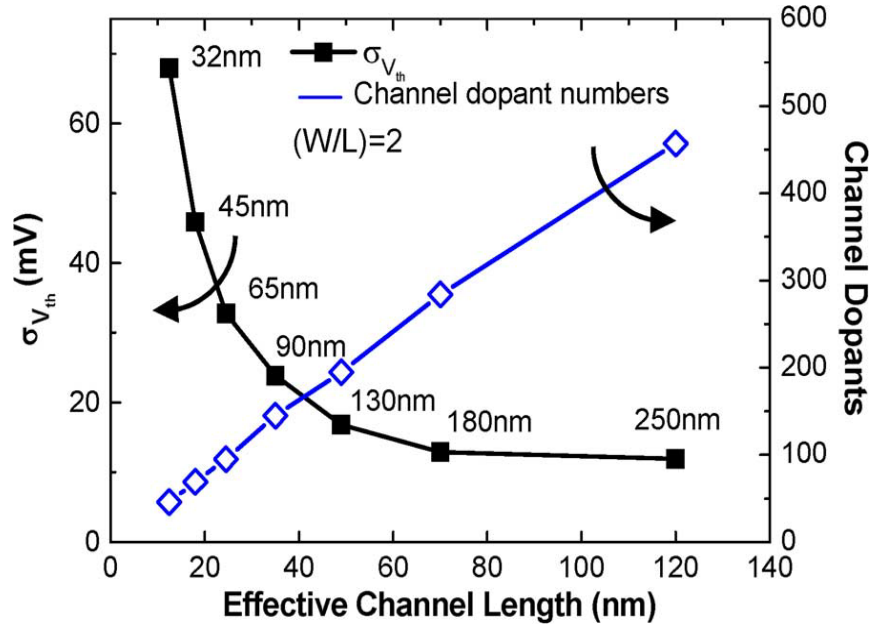


Figure 1.6 RDF induced σV_{th} scaling trend for sub-100nm MOS devices. The corresponding channel dopant number is also presented [59, 61, 67].

For instance, if a transistor has total of N number of dopants in its depletion, according to Poisson's distribution, the standard deviation of channel dopant from device to device follows $\sigma N = \sqrt{N}$. When N is on the order of hundreds, $\sigma N/N$ can be substantial. Since V_{th} of MOSFET devices is directly related to the ionization of the channel dopant, any slight variation in the dopant number or small change in the dopant placement can cause a significant V_{th} fluctuation as shown in Fig. 1.6. Many simulation and measurement studies have been focusing on RDF effects [66], and an analytical formula have been concluded. It is demonstrated that RDF induced V_{th} fluctuations follows a Gaussian distribution with standard deviation follows equation 1.1.

$$\sigma V_{th} = \frac{qT_{ox}}{\epsilon_{ox}} \sqrt{\frac{NW_d}{4L_{eff}W_{eff}}} \quad (1.1)$$

Where q is the electron charge, and ϵ_{ox} is oxide permittivity. N is the channel dopant concentration, and T_{ox} is the tunnel oxide thickness. W_{eff} and L_{eff} are the channel width and length for the MOSFET respectively, where W_d is the depletion layer width [68].

Equation 1.1 shows that V_{th} fluctuation due to RDF effect is inversely proportional to the square root of the device area, which is why it is especially sensitive in area constraint devices such as SRAM and aggressively scaled Flash. We traditionally rely on regular TCAD simulations with continuous doping profiles and compact models to quantify RDF in circuit analysis, but such methods become incorrect as the minimum feature size of a transistor is approaching the characteristic length of these atom-level effects. Instead, 3D Monte-Carlo atomistic simulations become necessary in order to achieve adequate accuracy [36]. Many modeling methods have been attempted, but today's computing power is still insufficient for

carrying out 3-D “atomistic” simulations on a large statistical scale [5]. Therefore, it is impossible to perfectly model the statistics of the RDF induced threshold voltage variation, let alone modeling the exact dopant placement with different device location combinations.

LER on the other hand, has caused little worry in the past since the critical dimensions of MOS devices were orders of magnitude larger than the edge roughness. However, it has been found that the percentage variations in transistor drive current and V_{th} caused by LER increases as device size shrinks, which are directly related to the short-channel effect (SCE) and drain-induced barrier lowering (DIBL) [34].

1.3 Impact of Process Variations on Memory Devices

1.3.1 Circuit and System Performance Degradation in SRAM

Many aspects of electronic manufacture and design are affected by the increase in memory device variability [69]. While process variation degrades performance and increases leakage in logic, its impact on SRAM and Flash is even stronger. High-density SRAM is ubiquitous for most modern ICs, and is one of the key design parameters used when technology advances. Its minimum feature size is constraint by the ever increasing requirement on the cache size, which generates significant leakage power [1]. Decrease in the supply voltage has exacerbated the variation impact on the SRAM functionality and manifested as reductions in static noise margins (SNM) and read current noise margin (SINM) [70]. These are the major challenges SRAM design faces.

1.3.2 Increased threshold voltage variations as Flash memory scales

Reference to Fig 1.3 (a), due to additional floating gate and control dielectric, Flash device fluctuations are more severe compared to conventional logic devices due

to the loss of gate control to the channel. The threshold voltage variation measured from the control gate is the floating gate variation divided by the coupling ratio, therefore resulting in a larger σV_{th} . The relationship of σV_{th} between logic device and Flash in the same technology node is depicted in equation 2, where α_{CP} is the coupling ratio [3].

$$\sigma \Delta V_{th_Flash} = \sigma \Delta V_{th_logic} / \alpha_{CP} \quad (1.2)$$

1.3.3 *Methods Employed to Mitigate the Impact of Process Variations*

The most critical challenge of variability increase is the development and utilization of the variability reduction methods. These methods range from pure process mitigation techniques, pure design mitigation techniques, and a combination of process-design mitigation techniques.

Pure process mitigation technique, by the name, refers to the improvement in the fabrication process in order to reduce the variability. For instance, RDF effect can be described by equation 1, which shows that decreasing in channel doping number and oxide thickness could theoretically reduce σV_{th} . However in reality, this improvement in σV_{th} as device scales down never happens. The reason is the formula never accounts for the increased gate leakage current due to thinner oxide and larger amount of oxide/interface defects, which is why industry decide to introduce high- κ dielectric and metal gate combination to help mitigating the impact of RDF and in hopes of enabling a return to the original scaling trend [7].

Pure design mitigation techniques, on the other hand, are trying to alleviate or remove the variability effect by using additional operations or different circuit/system designs without alter the fabrication technology itself. One of the design approaches for SRAM is using the dynamic forward body bias (FBB) to reduce leakage power,

where N-well is partially discharged 1 cycle before the word line (WL) by a programmable pulse [8]. Write and read assist circuits for SRAM is another example of mitigation method used to reduce the conflict between read and write operations in order to allow a lower supply voltage [71].

Combination process-design mitigation techniques are the ones that utilize the cooperation between process and design. One of the well-known methods in this category is when the industry decided to change the topology of SRAM from a “tall” design to a “wide” design [72]. This is due to the fact that the “wide” design provides better control on the critical dimensions for the fabrication process by aligning the poly in a single direction, therefore eliminating the diffusion corners, and relaxing the patterning constraints [7].

1.4 Chapter Organization

This dissertation intends to give in-depth analysis on process variability of CMOS memory devices by looking through 3D atomistic simulations, new mitigation technique and application. Chapter 2 illustrates the impact of RDF on the functionality and energy performance of a prototype 22nm SRAM. Monte Carlo technique is used instead of traditional TCAD methods to simulate individual transistor responses. Mixed mode simulations are then incorporated for cell level characterization. Chapter 3 introduces a hybrid SRAM-DRAM cell with cross capacitors that does not only help with SINM improvement, but also allows multi-bit storage for better memory performance. Chapter 4 characterizes the unique Flash Physical Unclonable Function (FPUF), which utilizes intrinsic variations in Flash memory to achieve security applications. Additional work on probing the orbital levels of engineered fullerenic molecules from a NVM cell is summarized in Chapter 5. Finally, conclusions and suggestions for future work will be discussed in Chapter 6.

REFERENCES

- [1] M. H. Abu-Rahma and M. Anis, *Nanometer Variation-Tolerant SRAM: Circuits and Statistical Design for Yield*, Springer, 2012.
- [2] H. Iwai, "Roadmap for 22 nm and beyond", *Microelectrode. Eng.*, vol. 86, pp.1520 -1528 2009.
- [3] D. Burnett, J. Higman, A. Hoefler, C. B. Li, and P. Kuhn, "Variation in natural threshold voltage of NVM circuits due to dopant fluctuations and its impact on reliability", *Int. Electronic Device Meeting Tech. Dig. (IEDM)*, pp.529, 2002.
- [4] K. Qian and C. J. Spanos, "A comprehensive model of process variability for statistical timing optimization", *Proc. SPIE*, vol. 6925, pp.178 -182, 2008.
- [5] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 μm MOSFETs: a 3D "atomistic" simulation study", *IEEE Trans. Electron Devices*, vol. 45, pp.2505 -2513, 1998.
- [6] B. Stine, D. Boning, and J. Chung, "Analysis and decomposition of spatial variation in integrated circuit processes and devices", *IEEE Trans. Semicond. Manuf.*, vol. 10, no. 1, pp.24 -41, 1997.
- [7] K. Kuhn, C. Kenyon, A. Kornfe K. Kuhn , C. Kenyon , A. Kornfeld , M. Liu , A. Maheshwari , W.-K. Shih, S. Siva Kumar, G. Taylor, P. VanDerVoorn and K. Zawadzki "Managing process variation in Intel's 45-nm CMOS technology", *Intel Tech. J.*, vol. 12, no. 2, pp.93 -110, 2008.
- [8] X. Liang and D. Brooks, "Mitigating the impact of process variations on CPU register file and execution units", *Proc. Int. Symp. Microarchitecture*, 2006.
- [9] B. Gassend, D. Clarke, M. van Dijk, and S. Devadas, "Silicon physical random functions", *Proc. Computer Communication Security Conf.*, pp.148 -160, 2002.
- [10] K. Agarwal and S. Nassif, "Characterizing process variation in nanometer CMOS", *Proc. IEEE/ACM DAC*, pp.396 -399, 2007.

- [11] L.-T. Pang and B. Nikoli, "Impact of layout on 90 nm CMOS process parameter fluctuations", *Symp. VLSI Circuits Dig. Tech. Papers*, pp.69 -70, 2006.
- [12] J. P. Cain and C. J. Spanos, "Electrical linewidth metrology for systematic CD variation characterization and causal analysis", *Proc. SPIE Int. Soc. Opt. Eng.*, 2003.
- [13] Y. Borodovsky, "Impact of local partial coherence variations on exposure tool performance", *SPIE*, vol. 2440, pp. 750-770, 1995.
- [14] C. Hedlund, H. Blom and S. Berg, "Microloading effect in reactive ion etching," *J. of vacuum science and tech.*, vol. 12, pp. 1962–1965, 1994.
- [15] S. Saha, "Modeling Process Variability in Scaled CMOS Technology," *Design Test of Computers*, IEEE, vol. PP, no. 99, 2010.
- [16] V. Moroz, L. Smith, X.-W. Lin, D. Pramanik and G. Rollins, "Stress-aware design methodology", *Intl. Symp. Quality Elec. Design*, 2006.
- [17] L.W. Liebmann et al., "TCAD development for lithography resolution enhancement", *IBM J. of Res. and Dev.*, vol. 45, pp. 651-665, Sep 2001. [18=14]
- [18] J.-Y. Lai, N. Saka and J.-H. Chun, "Evolution of copper-oxide damascene structures in chemical mechanical polishing", *J. of Electrochem. Soc.*, pp. G31-G40, 2002.
- [19] J. Watts, N. Lu, C. Bittner, S. Grundon and J. Oppold, "Modeling FET variation within a chip as a function of circuit design and layout choices", *Nanotech Workshop on Compact Modeling*, pp. 87-92, 2005.
- [20] J. Tschanz, K. Bowman, and V. De, "Variation-tolerant circuits: circuit solutions and techniques", *Proceedings of the 42nd Annual Conference on Design automation*, pp. 762–763, 2005.

- [21] S. Borkar, T. Karnik, S. Narendra, J. Tschanz, A. Keshavarzi, and V. De, “Parameter variations and impact on circuits and microarchitecture”, *Proceedings of the 40th conference on Design automation*, pp. 338–342, 2003.
- [22] T. Karnik, S. Borkar, V. De, “Sub-90 nm technologies: challenges and opportunities for CAD”, *ICCAD '02: Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, pp. 203–206, 2002.
- [23] S. Borkar, T. Karnik, and V. De, “Design and reliability challenges in nanometer technologies”, *Proceedings of the 41st Annual Conference on Design Automation*, pp. 75–75, 2004.
- [24] H. Masuda, S. Ohkawa, A. Kurokawa, and M. Aoki, “Challenge: variability characterization and modeling for 65- to 90-nm processes”, *Proceedings of IEEE Custom Integrated Circuits Conference*, pp. 593–599, 2005.
- [25] J.A. Croon, W. Sansen, and H.E. Maes, *Matching Properties of Deep Sub-Micron MOS Transistors*, Springer, New York, 2005.
- [26] T.-C. Chen, “Where is CMOS going: trendy hype versus real technology”, *Proceedings of the International Solid-State Circuits Conference ISSCC*, pp. 22–28, 2006.
- [27] K. Kuhn, “CMOS transistor scaling past 32 nm and implications on variation”, *Advanced Semiconductor Manufacturing Conference (ASMC), IEEE/SEMI*, pp. 241–246, July 2010.
- [28] B. Wong, A. Mittal, Y. Cao, and G.W. Starr, *Nano-CMOS Circuit and Physical Design*, Wiley-Interscience, New York, 2004.
- [29] J.-T. Kong, “CAD for nanometer silicon design challenges and success”, *IEEE Trans. Very Large Scale Integr. Syst.*, pp. 1132–1147, 2004.

- [30] A. Srivastava, D. Sylvester, and D. Blaauw, *Statistical Analysis and Optimization for VLSI: Timing and Power (Series on Integrated Circuits and Systems)*, Springer, Boston, 2005.
- [31] P. Friedberg, Y. Cao, J. Cain, R. Wang, J. Rabaey, and C. Spanos, "Modeling within-die spatial correlation effects for process-design co-optimization", *Proceedings of the Sixth International Symposium on Quality of Electronic Design*, pp. 516–521, 2005.
- [32] K. Bernstein, "High-performance CMOS variability in the 65-nm regime and beyond", *IBM J. Research and Development*, vol. 50, nos. 4-5, pp. 433-449, 2006.
- [33] A. Asenov, "Origin of the Asymmetry in the Magnitude of the Statistical Variability of n- and p-Channel Poly-Si Gate Bulk MOSFETs", *IEEE Electron Device Letters*, vol. 29, no. 8, pp. 913-915, 2008.
- [34] Y. Cheng, C. Hu, *MOSFET Modeling and BSIM User Guide*, Kluwer Academic Publishers, Boston, 1999.
- [35] A. I. Chou, et al., "Modeling of stress-induced leakage current in ultrathin oxides with the trap-assisted tunneling mechanism", *Appl. Phys. Lett.*, vol. 70, pp.3407 -3409, 1997.
- [36] S. Kamohara, D. Park, and C. Hu, "Deep-trap SILC (Stress-Induced Leakage Current) model for nominal and weak oxides", *Proc. IRPS*, pp.57 -61, 1998.
- [37] D. J. Di Maria, and E. Cartier, "Mechanism for stress-induced leakage currents in thin silicon dioxide films", *Journal of Applied Physics*, vol. 78, no. 6, pp. 3883-3894, 1995.
- [38] N.Naruke, S. Taguchi, and M. Wada, "Stress Induced Leakage Current Limiting to Scale Down EEPROM Tunnel Oxide", *IEDM Technical Digest*, p. 424, 1988.

- [39] J. De Blauwe, J. Van Houdt, D. Wellekens, G. Groeseneken, and H. E. Maes, "SILC Related Effects in Flash E²PROM's- PartI: A Quantitative Model for Steady-State SILC", *IEEE Trans.on Electron Devices*, vol. 45, no. 8, pp. 1745-1750, 1998.
- [40] C. E. Blat, E. H. Nicollian, and E. H. Poindexter, "Mechanism of negative-bias-temperature instability", *J. Appl. Phys.*, vol. 69, p. 1712, 1991.
- [41] H. Ushizaka and Y. Sato, "The process dependence on positive bias temperature aging instability of p+(B) polysilicon-gate MOS devices", *IEEE Trans. Electron Devices*, vol. 40, p. 932, 1993.
- [42] J. F. Zhang and W. Eccleston, "Positive bias temperature instability in MOSFETs", *IEEE Trans. Electron Devices*, vol. 45, pp.116 -124, 1998.
- [43] B. S. Doyle, B. J. Fishbein, and K. R. Mistry, "NBTI-enhanced hot carrier damage in p-channel MOSFET's", *Proc. IEDM*, p. 529, 1991.
- [44] C. Hu, S. C. Tam, F. C. Hsu, P. K. Ko, T. Y. Chan, and K. W. Terrill, "Hot-electron-induced MOSFET degradation model, monitor, and improvement", *IEEE J. Solid-State Circuits*, vol. SC-20, p. 295, 1985.
- [45] D. J. DiMaria and J. W. Stasiak, "Trap creation in silicon dioxide produced by hot electrons", *J. Appl. Phys.*, vol. 65, p. 2342, 1989.
- [46] P. Heremans, R. Bellens, G. Groeseneken, and H. E. Maes, "Consistent model for the hot-carrier degradation in n-channel and p-channel MOSFET's", *IEEE Trans. Electron Devices*, vol. 35, p. 2194, 1988.
- [47] Y. Miura and Y. Matukura, "Investigation of Silicon-Silicon Dioxide Interface Using MOS Structure", *Jap.J.Appl.Phys.*, vol. 5, pp. 180, 1966
- [48] R. Moonen, "Study of time-dependent dielectric breakdown on gate oxide capacitors at high temperature", *Proc. of 14th IPFA*, p. 288–91, 2007.

- [49] K. Takeuchi, T. Nagumo, K. Takeda, S. Asayama, S. Yokogawa, K. Imai, and Y. Hayashi, "Direct observation of RTN-induced SRAM failure by accelerated testing and its application to product reliability assessment", *Symposium on VLSI Technology (VLSIT)*, pp. 189–190, June 2010.
- [50] N. Tega, H. Miki, F. Pagette, D. Frank, A. Ray, M. Rooks, W. Haensch, and K. Torii, "Increasing threshold voltage variation due to random telegraph noise in FETs as gate lengths scale to 20 nm", *Symposium on VLSI Technology*, pp. 50–51, June 2009.
- [51] B. Razavi, *Design of Analog CMOS Integrated Circuits*, McGraw-Hill, New York, 2000.
- [52] S. O. Toh, Y. Tsukamoto, G. Zheng, L. Jones, L. Tsu-Jae King, and B. Nikolic, "Impact of random telegraph signals on V_{min} in 45nm SRAM", *IEEE International Electron Devices Meeting*, Baltimore, pp. 767-770, 2009.
- [53] N. Tega, H. Miki, T. Osabe, A. Kotabe, K. Otsuga, H. Kurata, S. Kamohara, K. Tokami, Y. Ikeda, and R. Yamada, "Anomalously Large Threshold Voltage Fluctuation by Complex Random Telegraph Signal in Floating Gate Flash Memory", *International Electron Devices Meeting*, San Francisco, pp. 491-494, 2006.
- [54] M. J. Uren, D. J. Day, and M. J. Kirton, "1/F and Random Telegraph Noise in Silicon Metal-Oxide-Semiconductor Field-Effect Transistors", *Appl. Phys. Lett.*, vol.47, pp. 1195-1197, 1985.
- [55] S. Machlup, "Noise in semiconductors: spectrum of a two-parameter random signal", *J. Appl. Phys.*, vol. 25, pp.341-343, 1954.
- [56] K. Abe , A. Teramoto , S. Sugawa and T. Ohmi "Understanding of traps causing random telegraph noise based on experimentally extracted time constants and amplitude", *Proc. IRPS*, pp.381 -386, 2011.

- [57] H.-J. Cho, Y. Son, B.-C. Oh, S. Lee, J. H. Lee, B.-G. Park and H. Shin, "Study on time constants of random telegraph noise in gate leakage current through hot carrier stress test", *IEEE Electron Device Lett.*, vol. 31, no. 9, pp.1029 - 1031, 2010.
- [58] C. Mukherjee and C. K. Maiti, *Nanowires-Recent Advances*, Intech, 2012.
- [59] "International Technology Roadmap for Semiconductors," 2012.
- [60] S. V. Kumar, C. H. Kim and S. S. Sapatnekar, "An analytical model for negative bias temperature instability", *Proc. Int. Conf. Comput.-Aided Des.*, pp.493 -496, 2006.
- [61] Yun Ye, Frank Liu, Min Chen, Sani Nassif, and Yu Cao, "Statistical Modeling and Simulation of Threshold Variation Under Random Dopant Fluctuations and Line-Edge Roughness", *IEEE Transaction on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 19(6), pp. 987-996, 2011.
- [62] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of intrinsic parameter fluctuations in decananometer and nanometer-scale MOSFETs", *IEEE Trans. Electron Devices*, vol. 50, no. 9, pp. 1837–1852, Sep. 2003.
- [63] B. Cheng, S. Roy, G. Roy, F. Adamu-Lema, and A. Asenov, "Impact of intrinsic parameter fluctuations in decanano MOSFETs on yield and functionality of SRAM cells", *Elsevier Solid-State Electron.*, vol. 49, pp. 740–746, 2005.
- [64] K. J. Kuhn, "Reducing Variation in Advanced Logic Technologies: Approaches to Process and Design for Manufacturability of Nanoscale CMOS", *IEDM Technical Digest*, pp. 471–474. , December 2007.
- [65] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of Fundamental Threshold Voltage Variations for High-Density SRAM and Logic Circuits", *Symp. on VLSI Technology*, pp. 15-16, 1994.

- [66] T. Mizuno, J. Okamura, and A. Toriumi, "Experimental study of threshold voltage fluctuation due to statistical variation of channel dopant number in MOSFETs", *IEEE Trans. Electron Devices*, vol. 41, pp.2216 -2221, 1994.
- [67] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45 nm design exploration", *IEEE Trans. Electron Devices*, vol.53, no. 11, pp. 2816–2823, Nov. 2006.
- [68] H.-S. Wong and Y. Taur, "Three-dimensional 'atomistic', simulation of discrete microscopic random dopant distributions effects in sub-0.1 μm MOSFETs", *IEDM Tech. Dig.*, pp.705 -708, 1993.
- [69] S. Saxena, C. Hess and H. Karbasi, "Variation in transistor performance and leakage in nanometer-scale technologies", *IEEE Trans. Electron Devices*, vol. 55, pp.131 -144, 2008.
- [70] K. Agarwal, S. Nassif, "Statistical analysis of SRAM cell stability", *Proceedings of the 43rd Annual Conference on Design Automation*, pp. 57–62, 2006.
- [71] K. Kuhn, "CMOS transistor scaling past 32 nm and implications on variation", *IEEE/SEMI*, pp. 241–246, July 2010.
- [72] H.-W. Kim, J.-Y. Lee, J. Shin, S.-G. Woo, H.-K. Cho and J.-T. Moon, "Experimental investigation of the effect of LWR on sub-100-nm device performance", *IEEE Transactions on Electron Devices*, Vol. 51, Issue 12, pp.1984–1988, December 2004.

CHAPTER 2

IMPACT OF RDF ON 22NM SRAM NOISE MARGINS

2.1 *Abstract*

Impact of RDF on the static noise margin (SNM) and read current margin (SINM) of a prototype 22nm six- transistor (6T) SRAM was investigated using TCAD modeling. Individual device statistics of threshold voltages (V_{th}) and transport related parameters were first extracted for NFETs and PFETs. SNM and SINM characteristics of the corresponding SRAM cells were then analyzed. Two methods to emulate the impact of RDF were simulated — modulating gate work function, and uniform scaling of the continuum dopant distribution; compared to RDF devices, both methods underestimate V_{th} and SNM variations. SNM and SINM tail distributions are also analyzed for design considerations.

2.2 *Introduction*

Driven by the improvements on performance and cost of today's integrated circuits, new generations of SRAM cells are being aggressively scaled down to 22nm technology node and beyond [1]. Continued advancement in the CMOS technologies reduce the feature sizes closer to atomic dimensions, lowering supply voltage and power consumption. Cells and systems based on such devices are becoming more susceptible to variations and mismatches, causing various scaling challenges [2]. One of the most pronounced scaling effect is the V_{th} variations caused by random dopant fluctuation (RDF) in the channel region [3, 4].

RDF effect, as mentioned in Chapter 1, refers to the microscopic variations in the discrete number and arrangement of the channel dopant as device feature shrinks down dramatically [5, 6]. The effect was first explored in the seventies [4], and

later recognized to be the major contributor to device variations at sub-100nm dimensions [7, 8]. Since RDF is entirely intrinsic and cannot be eliminated through careful control of the fabrication process, it has become the functionality bottleneck for minimum-feature device, such as area-constrained SRAM cells [9].

In this study, dopant fluctuations were introduced into a prototype 22nm 6T SRAM cell using Monte Carlo techniques. The statistics of threshold voltages and transport related parameters of individual devices were first extracted and presented in section 2.3. SNM and SINM characteristics of the corresponding SRAM cells were then analyzed using mixed mode simulation and summarized in section 2.4. Two alternative approaches—gate work function modulation, and uniform scaling of the continuum dopant distribution—were also studied to assess how well they can emulate the impact of RDF.

2.3 Individual Transistor Simulations

2.3.1 Simulation and Analysis Methodologies

Fig.2.1 illustrates the cross section of a prototype 22nm planar device employed in this study, with an instance of RDF doping profile. Statistical characteristics were analyzed individually for Pull-down (PD) NFET, Pull-up (PU) PFET, and Pass-gate (PG) NFET. Process simulations were performed in TSUPREM-4 [10] using a prototype process flow to generate the nominal device structures. RDF was then introduced into the nominal structure with Monte Carlo simulated doping profiles assuming Poisson distributions for the dopants.

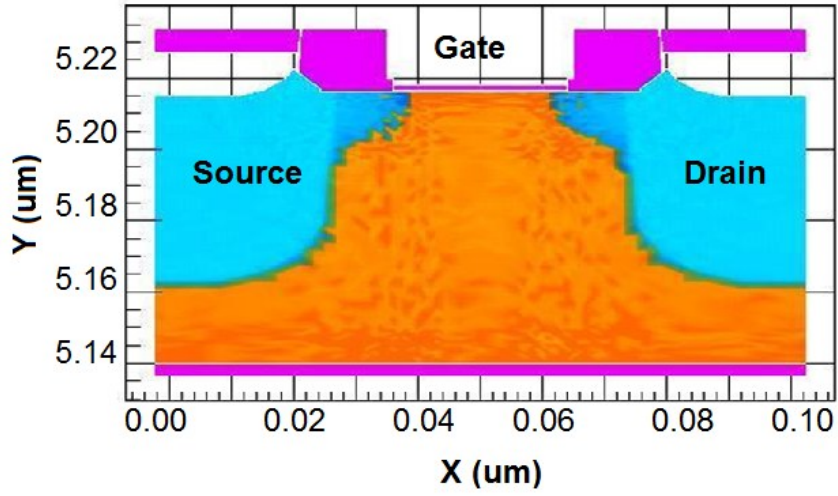


Figure 2.1 Cross section of the prototype 22nm NFET device with an instance of Monte Carlo introduced discrete random dopant fluctuation in the channel.

2.3.2 Threshold Voltage Distributions

The resulting linear and saturation threshold voltage distributions of PD NFET are shown in Fig.2.2 as an example. The structure parameters and the corresponding

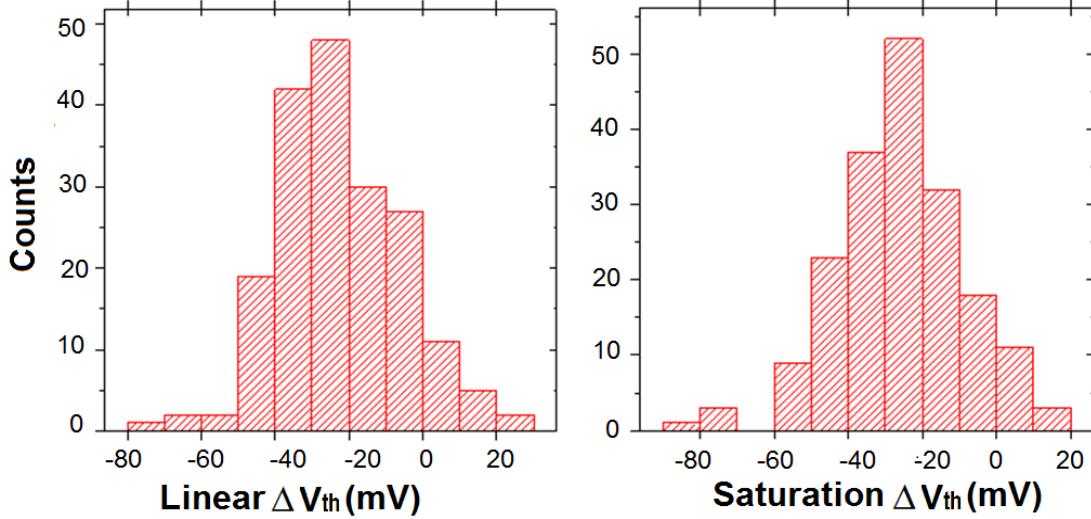


Figure 2.2 a) linear threshold voltage distribution b) saturation threshold voltage distribution of RDF Pull-down NFETs referenced to nominal device results.

statistical V_{th} results of all three types of devices are summarized in Table 2.1. As device width shrinks down, the amount of channel dopant drops, making the device-to-device fluctuations in discrete doping profile more prominent. Therefore, PU PFET, with the smallest geometries and dopant number, shows the largest threshold voltage variation (σV_{th}) among the three.

Table 2.1 Standard deviations of threshold voltages for RDF devices at both linear (σV_{tlin}) and saturation (σV_{tsat}) regions using a sample size of 200.

Device	Width	σV_{tlin} (mV)	σV_{tsat} (mV)
Pull Down NFET	8L	16.6	17.6
Pass Gate NFET	5L	17.5	17.6
Pull Up PFET	3L	28.4	36.7

Two alternative approaches were investigated to understand how well the effects of RDF could be emulated without using discrete doping profiles. Gate work function modulation (WF) and continuum channel dopant scaling (Continuum) were performed. Threshold voltage variations were extracted for all three devices using both methods. Gate work function scaling captures σV_{th} of RDF effect by varying the gate work function of the nominal device, so that the resulting threshold voltage range is the same as that of the simulated RDF devices. This is the most intuitive way to imitate the threshold voltage variations in RDF without considering the doping number or location in the channel. Scaling the channel dopant (Continuum), on the other hand, accounts for the differences in the number of channel dopants; however, it does not consider the discreteness of the dopant or the random dopant arrangement in space.

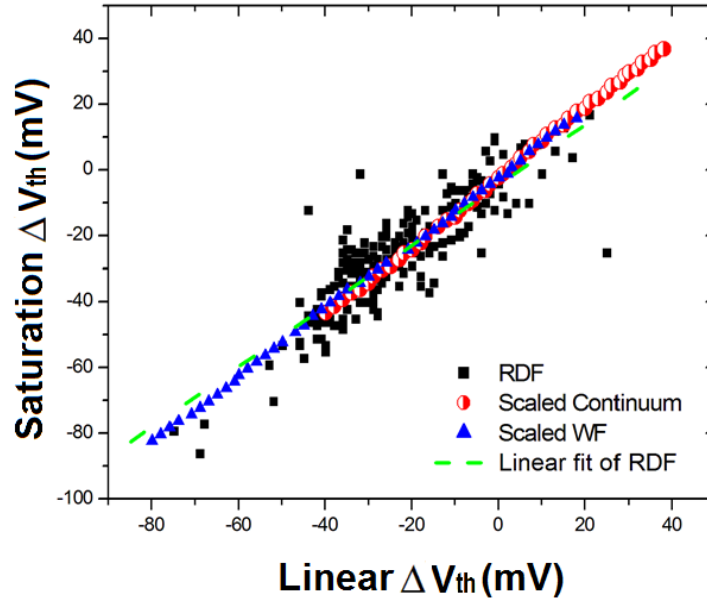


Figure 2.3 Saturation ΔV_{th} vs linear ΔV_{th} referenced to the nominal device values for all three devices: RDF, Continuum and WF. An average V_{th} shift was observed between RDF devices and Continuum devices.

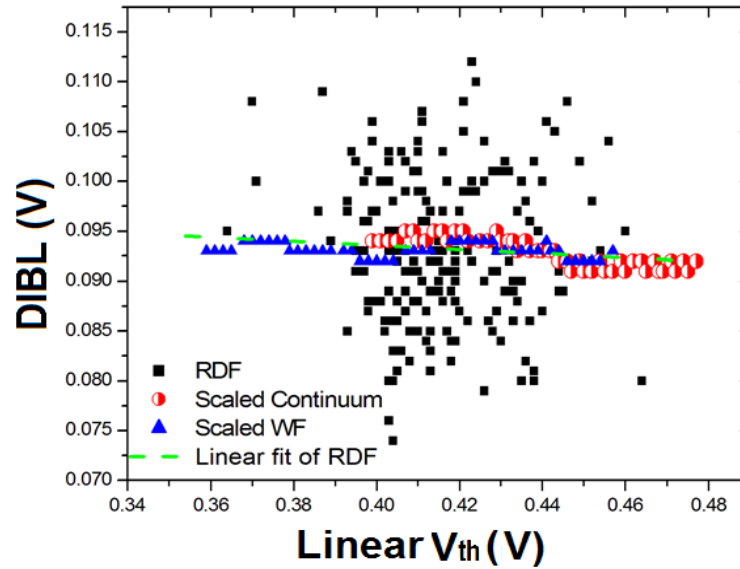


Figure 2.4 DIBL vs linear V_{th} for all three cases. DIBL of WF devices only showed 1mV fluctuations which was caused by the granularity of data extraction.

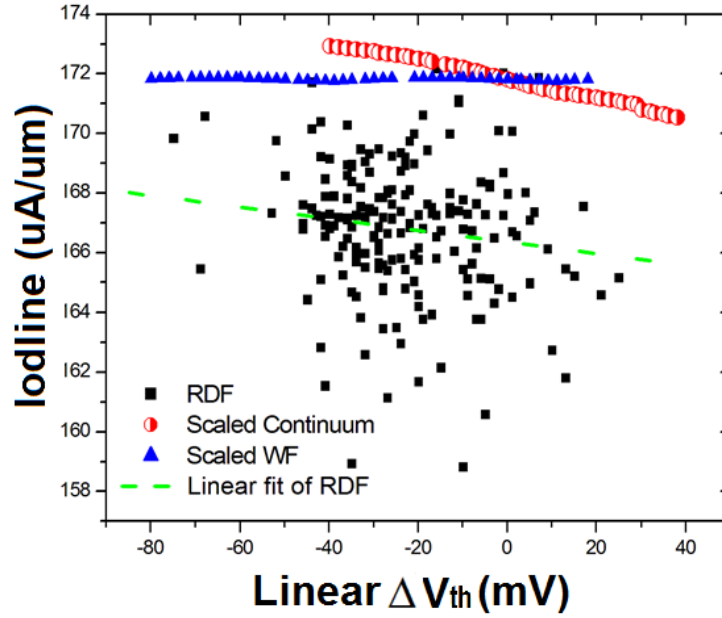


Figure 2.5 Linear overdrive current vs linear ΔV_{th} . WF devices showed no V_{th} dependency due to unvaried channel doping profile.

2.3.3 Transport Characteristics

The linear/saturation V_{th} and carrier transport related properties such as drain-induced barrier lowering (DIBL) and overdrive current (I_{odlin}) were summarized in Fig. 2.3, 2.4 and 2.5 respectively. Since DIBL and on resistance (R_{on}) are affected by doping variations, Continuum devices captured DIBL/ R_{on} better than WF devices.

In addition, a shift in V_{th} exists between the Continuum and RDF devices, which is evident in Fig. 2.3. This phenomenon was first observed in [7], where the average shift of the threshold voltage was attributed to the inhomogeneity of channel potential due to the randomness and discreteness of the channel dopants. In Fig. 2.6, an instance of potential barrier vs channel location was graphed out for the nominal and RDF devices having the same number of channel dopants at threshold voltage bias. It is clear to see that due to the random dopant distribution, the potential barrier

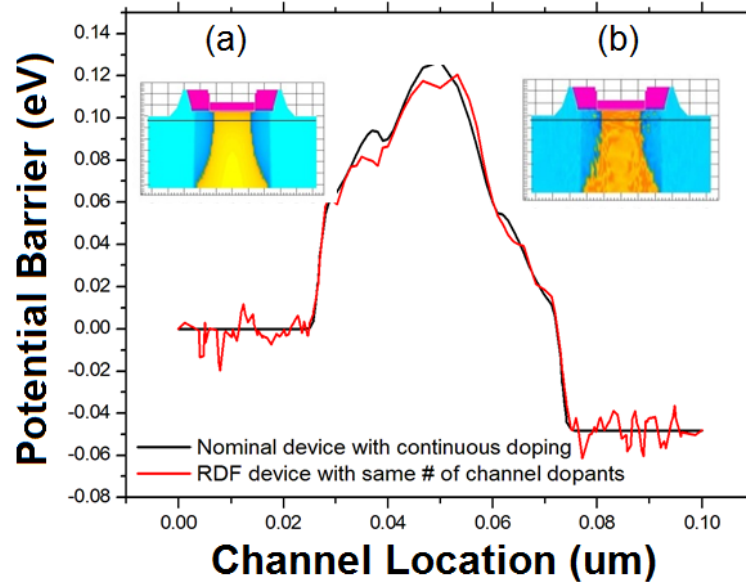


Figure 2.6 Potential barrier at threshold voltage bias for nominal and RDF devices having the same number of channel dopant. Inserts: example of a a) nominal and b) RDF device doping profile.

profile of the RDF device fluctuates along the channel region, causing the maximum barrier height to be lower than that of the Continuum device, in spite of the fact that they have the same number of channel dopants.

2.4 *Mixed Mode SRAM Simulation*

2.4.1 *Mixed Mode Simulation Methodologies*

The stability and reliability of SRAM cells during read and write operations are often characterized by the noise margins that need to be maintained [9, 11]. The benchmarks for accessing the SRAM cell stability are usually the static noise margin (SNM) and current noise margin (SINM) during read operation [4, 5, 6], during which the cell states are most vulnerable to external signal perturbations. SNM is the

maximum tolerable DC noise voltage at a storage node without causing a read disturbance, while SINM is determined from an N-curve measurement [12].

In the mixed mode simulation, RDF, WF and Continuum devices are used individually to create their corresponding SRAM groups. For RDF SRAMs, each of PD, PG and PU contributes 200 devices into the device selection pool with a naturally Gaussian distribution in V_{th} due to the Monte Carlo generated doping profiles, and the devices are then randomly selected to form the desired SRAM cells. The WF device pool was created by using the nominal device threshold voltage as the average V_{th} , and the RDF σV_{th} was used to calculate the range of work function required. The WF SRAM cells were then randomly selected from the device pool with the same Gaussian probability seen in the RDF devices. Instead of matching the σV_{th} , the Continuum devices used the nominal device doping profile as the average device doping profile, and the RDF dopant distribution to obtain the range of amount of channel dopant required. Again, the dopant numbers follow the same distribution as the RDF case.

2.4.2 *Static Noise Margin (SNM) and Read Current Noise Margin (SINM)*

22nm 6T SRAMs were then set up using RDF, Continuum, or WF devices. The SNM and SINM were simulated using sample size of 1000; statistic results are illustrated in Fig. 2.7. It is interesting to see that, in spite of the fact that only Continuum devices were able to capture DIBL and Ron variations, SRAMs built using Continuum and WF devices showed similar distribution in SNM and SINM; however, neither of the simpler methods predicted the same results as that of the RDF cells.

In addition, since there is an average threshold voltage shift between the mean RDF and the nominal device, the comparison between RDF and the other two cases may be distorted. In order to eliminate the contribution from this average V_{th} shift

and focus only on the RDF effect, a new group of SRAM cells having the same average V_{th} as RDF was simulated. Since previously WF and Continuum sets are seen to show similar results in SNM and SINM variations, shifted WF devices were chosen over shifted channel doping devices to investigate this effect for a simpler simulation scheme. The results are shown in Fig. 2.8. With lower V_{th} , shifted WF devices showed a slightly smaller SNM and larger SINM; however, σ_{SNM} and σ_{SINM} remained the same as the original WF SRAMs. The results indicate that shifting the average threshold voltage without considering the dopant distribution cannot recreate the RDF results. Standard deviation of the standard deviation (SoS) is used to evaluate the difference in variations among the four cases to assess if they are significant enough to be seen as statistically different. Shifted WF SRAMs, despite having the same average V_{th} and σV_{th} as the RDF devices, still have a smaller standard deviation. The SoS confirms that the differences among of the four SRAM sets are significant (over 3σ), as seen in Table 2.2.

Table 2.2 a) SNM and b) SINM variations for SRAMs composed of RDF, Continuum, WF and shifted WF devices.

(a)		
SRAM device	σ_{SNM} (mV)	SoS of SNM (mV)
RDF	10	Approximately 0.20
Continuum	8.6	
WF	8.7	
Shifted WF	8.5	

(b)		
SRAM device	σ_{SINM} (μA)	SoS of SINM (μA)
RDF	5.6	Approximately 0.10
Continuum	5	
WF	5.1	
Shifted WF	5.22	

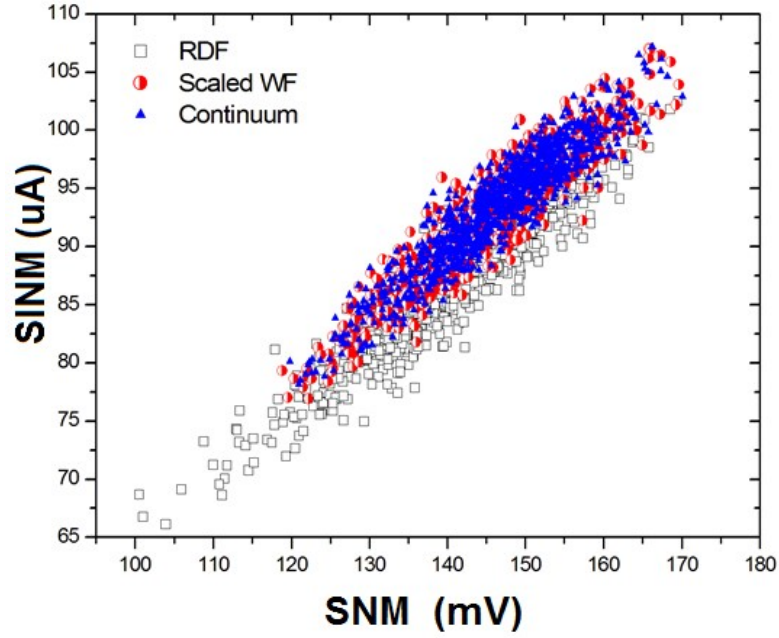


Figure 2.7 SRAMs composed of WF and Continuum devices show similar SNM and SINM, and both are larger than RDF cells.

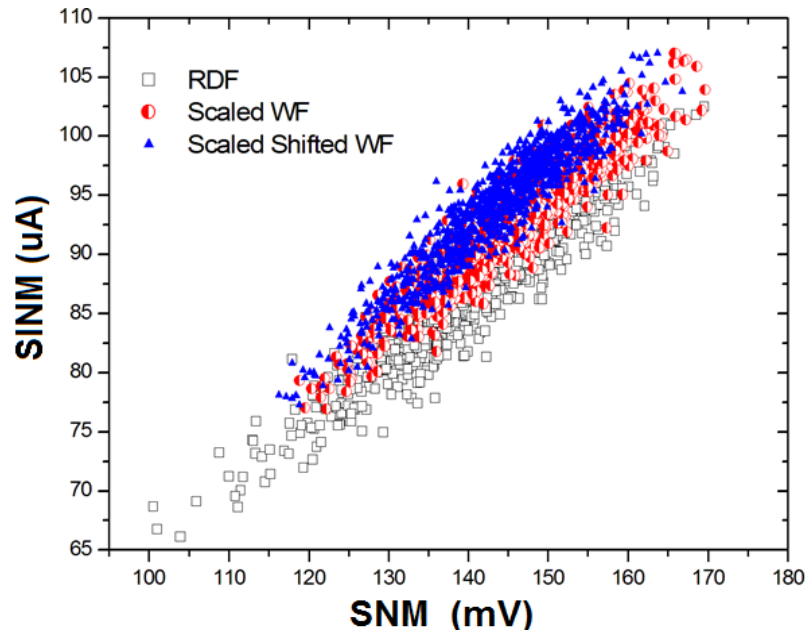


Figure 2.8 SNM and SINM for SRAMs composed of RDF, WF and Shifted WF devices.

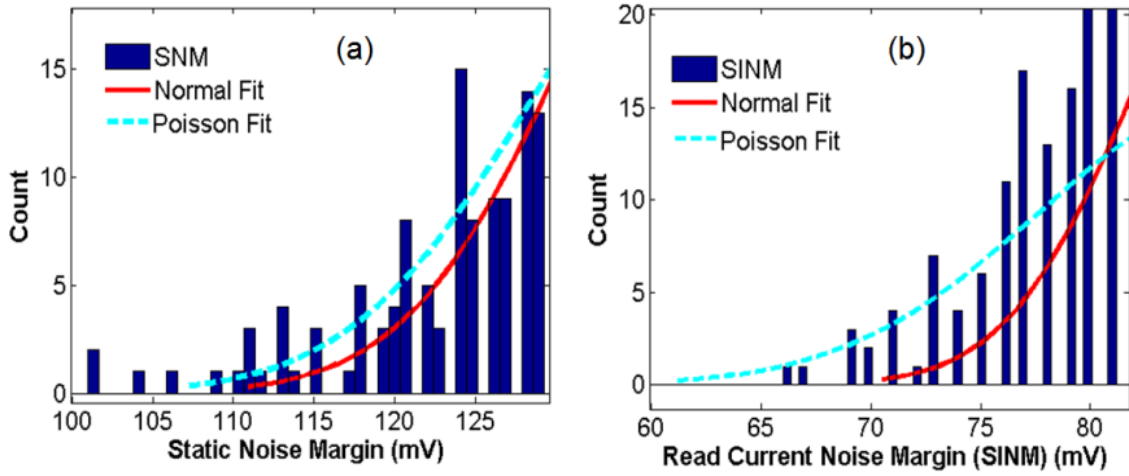


Figure 2.9 Lower tail distributions for (a) SNM and (b) SINM of 22nm prototype SRAM with both normal and Poisson curve fittings.

2.4.3 Tail Distributions of SNM and SINM for Design Considerations

Tail distributions of SNM and SINM are plotted in Fig. 2.9 (a) and (b) respectively. It is clear that due to the discrete nature of the RDF effect, the tail distributions are more accurately modeled through Poisson distributions rather than normal distributions. The results are much longer lower tails, which are critical for circuit and system designer. This is because these extreme cases set the lower bound for the device performance, and is becoming a limiting factor for yield.

2.5 Conclusion

TCAD modeling of RDF impact on a prototype 22nm 6T SRAM is presented in this paper. Statistics of V_{th} variations and transport related parameters were obtained for individual types of devices. Two methods to emulate the RDF effect were investigated—scaling of the continuum dopant distribution, and gate work function modulation. Since the Continuum devices contained some channel doping information,

they were able to capture DIBL and Ron better than WF devices, however, neither could reproduce the RDF effect adequately. Mixed mode simulations were then employed to simulated 6T SRAM cells composed of three different types of devices, and the corresponding SNM and SINM characteristics were extracted. Continuum and WF devices showed similar SNM/ σ SNM and SINM/ σ SINM, in spite of the fact that Continuum devices were able to deliver DIBL & Ron variations. From a statistical point of view, in order to accurately examine the effect of random dopant fluctuations within the 22nm SRAM cell, full scale Monte Carlo simulated doping profile is necessary. WF and Continuum devices are only capable of showing the general trend, but neither method was seen to emulate the RDF effect satisfactorily.

REFERENCES

- [1] International Technology Roadmap for Semiconductors (ITRS), 2011.
- [2] K.J. Kuhn, "Reducing variation in advanced logic technologies: Approaches to process and design for manufacturability of nanoscale CMOS," *IEDM Tech. Dig.*, pp. 471–474, Dec. 2007.
- [3] R. W. Keyes, "The effect of randomness in the distribution of impurity atoms on FET thresholds," *Appl. Phys.*, vol. 8, pp. 251–259, 1975.
- [4] D. J. Frank, Y. Taur, M. Jeong, and H.-S. P. Wong, "Monte Carlo modeling of threshold variation due to dopant fluctuations", *Symp. VLSI Technol. Dig.*, pp.169 -170 1999.
- [5] X. Tang, V. De and J. Meindl "Intrinsic MOSFET parameter fluctuations due to random dopant placement", *IEEE Trans. VLSI Syst.*, vol. 5, pp.369 1997.
- [6] B. Cheng , S. Roy and A. Asenov "The impact of random doping effects on CMOS SRAM cell", *Proc. Eur. Solid-State Circuits Conf.*, pp.219 2004.
- [7] H. S. Wong and Y. Taur, "Three-dimensional atomistic simulation of discrete random dopant distribution effects in sub-0.1 μ m MOSFET's", *IEDM Tech. Dig.*, pp. 705, 1993.
- [8] A. Asenov "Random dopant induced threshold voltage lowering and fluctuations in sub", *IEEE Trans. Electron Devices*, vol. 45, pp.2505 1998.
- [9] A. Bhavnagarwala , X. Tang and J. Meindl "The impact of intrinsic device fluctuations on CMOS SRAM cell stability", *IEEE J. Solid-State Circuits*, vol. 36, pp.658 2001.
- [10] Synopsys, Inc., TSUPREM-4, Two-Dimensional Device Simulation Program Version Version D-2-1-.03 User's Manual
- [11] D. Burnett, K. Erington, C. Subramanian, and K. Baker, "Implications of fundamental threshold voltage variations for high-density SRAM and logic

circuits,” *Proc. Symp. VLSI Tech.*, pp. 15–16, June 1994.

- [12] E. Grossar , M. Stucchi , K. Maex and W. Dehaene "Read stability and write-ability analysis of SRAM cells for nanometer technologies", *IEEE J. Solid-State Circuits*, vol. 41, pp.2577, 2006.

CHAPTER 3

HYBRID SRAM-DRAM WITH CROSS CAPACITORS FOR MULTI-BIT STORAGE AND DISTURB STABILIZATION

3.1 *Abstract*

The hybrid SRAM-DRAM memory cell with cross capacitors is investigated for multi-bit storage and disturb stabilization capabilities. The design enables nanosecond internal context switching between SRAM and DRAM without accessing the external bus, thus improving data bandwidth and power consumption. The cross capacitor is also naturally differential and can stabilize SRAM read disturb when the same information is stored between SRAM and DRAM. How the two functions (storing multiple bits and perturb stabilization) can be both assigned and realized during operations is briefly discussed. Limitation of using conventional noise margin metric for the hybrid cell is characterized. Both N-curve and transient current noise margins are employed to evaluate the disturb immunity.

3.2 *Introduction and Motivation*

SRAM and DRAM are the predominant technologies used to implement memory in computer systems. Aggressively scaled SRAM cells dissipate considerable amount of static power [1, 2]. On the other hand, DRAM, although slower in access speed, can have higher density and more efficient energy consumption even considering refreshing operations. Integrating SRAM and DRAM into hybrid cells can potentially achieve faster access time, smaller area and less power [2, 3]. As feature size shrinks down into nanometer regime, device variability also becomes more severe, which significantly reduces SRAM reliability. In addition, in order to lower power consumption, the supply voltage needs to be scaled further, making SRAM

more vulnerable to various disturbs and mismatches. In this paper, we investigate a hybrid SRAM-DRAM cell [4] that allows nanosecond internal context switching between SRAM and DRAM branches with retention time longer than 10's milliseconds. The system advantages of the large hybrid memory bank were investigated in [4], and here we present the in-depth circuit analysis from 65nm technology. Scaling effect is also discussed through comparison with other technology cells. Having a DRAM cross capacitor that is inherently differential between the complementary bit lines does not only allows the hybrid cell to store multiple bits, but can also be used to stabilize read disturb and soft errors. Read current noise margin is used as the read stability metric for the hybrid cell, and is evaluated through both the conventional N-curve method [5] and transient current noise analysis.

3.3 *Hybrid Cell Structure*

Fig. 3.1 (a) and (b) illustrate two types of hybrid SRAM-DRAM cells. A typical six-transistor (6T) SRAM cell is augmented with either N pairs of traditional 1T1C DRAM cells, or equivalently N branches of 2T-1C DRAM cells across the Q and Qb nodes. The latter is preferred design that does not only save half of the required capacitances due to miller effect, but also eliminates additional capacitor mismatches. Each 2T1C DRAM branch consists of one cross capacitor and two pass transistors controlled by the enable (EN) signal. This configuration allows one hybrid cell to store the same amount of data as N regular SRAM cells, which increases the effective bit density and decreases static leakage. An implementation example of GPU register files shows 38% area reduction and 68% energy savings when compared with conventional SRAM designs [4].

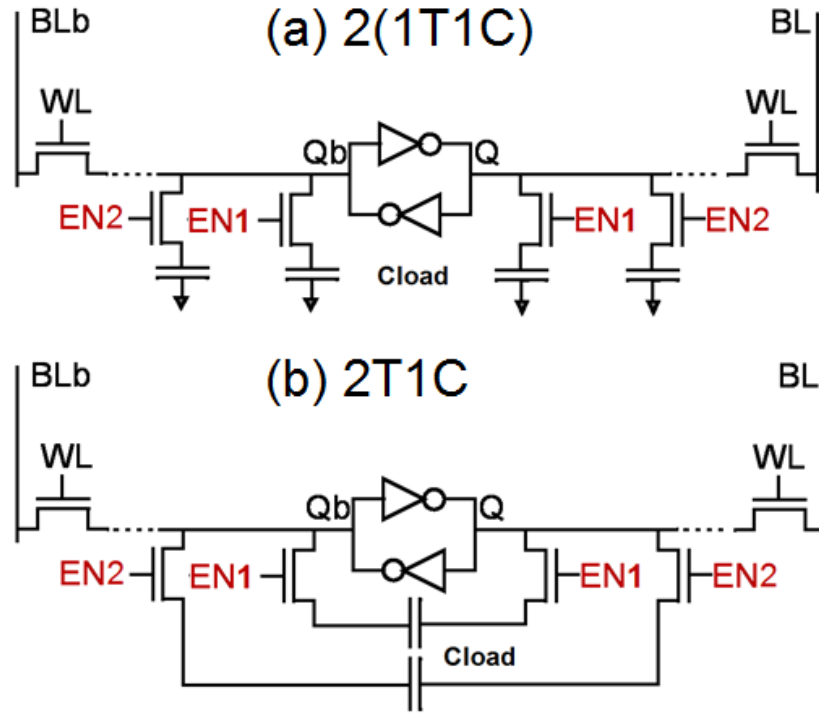


Figure 3.1 Schematics of the SRAM-DRAM hybrid cell with (a) 2 1T1c DRAM branches, and (b) 2T1C DRAM branches. N branches can be added to the cell with trade-off between stored bits and extra load on the internal storage nodes (Q and Qb). BL is bit line, BLb is complementary bit line, and WL is word line.

3.4 SRAM-DRAM Hybrid Cell Operations

3.4.1 Regular Operations

EN signals are turned off at regular operations to put DRAMs in retention. The cell then resembles a typical 6T SRAM cell bearing extra capacitive load at Q and Qb, which increases the SRAM access delay slightly, but improves the stability against disturbance and soft errors if one or more DRAM branches storing the same information as SRAM are turned ON. EN signals are turned off at regular operations

to put DRAMs in retention. The cell then resembles a typical 6T SRAM cell bearing extra capacitive load at Q and Qb, which increases the SRAM access delay slightly, but improves the stability against disturbance and soft errors if one or more DRAM branches storing the same information as SRAM are turned ON.

3.4.2 Context Switching between SRAM and DRAM

To manage active and dormant contexts between SRAM and DRAM branches, two new operations are introduced: STORE and RESTORE. STORE operation is used when the active SRAM context is stored to a specific DRAM node. The active context in SRAM is not affected and remains available. RESTORE operation is used when the dormant context in the i -th DRAM node is loaded to SRAM and being refreshed. The

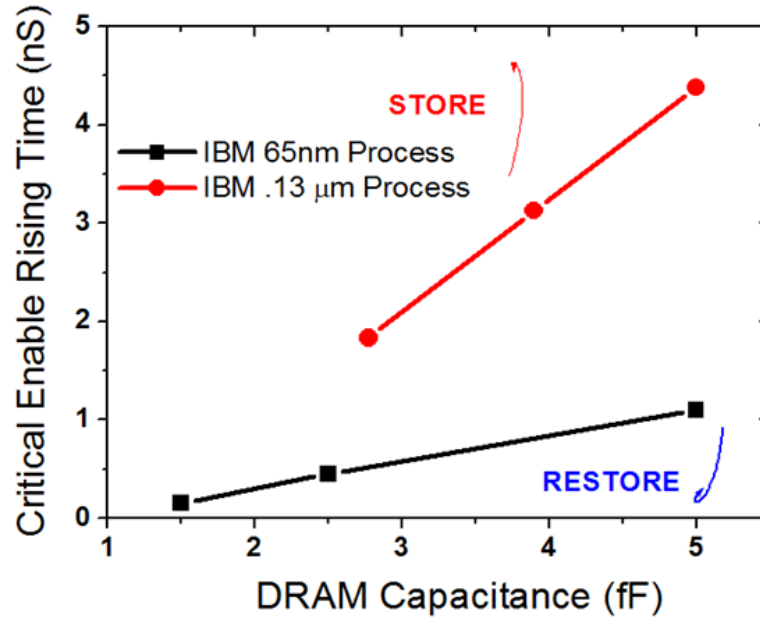


Figure 3.2 Critical EN rising time with various DRAM capacitances. STORE is performed when EN is slower than the critical time; otherwise, RESTORE is achieved.

EN_i is a ramping signal and its rise time determines whether the bit is stored to or restored from the i -th DRAM node. Fast rise time allows charges on the DRAM capacitor to quickly disturb the SRAM equilibrium and flip the active state if SRAM is storing different information as the DRAM node, resulting in a RESTORE operation; slow rise time of EN signal lets the DRAM capacitor to gradually charge or discharge during the STORE operation while maintaining the original SRAM state due to small and slower disturbance. Therefore, understanding the critical EN signal ramping speed is crucial for the hybrid cell operation.

Fig. 3.2 shows the critical EN rising time that separates the STORE and RESTORE, which is around a few nanoseconds and scales with the cross capacitance as well as technology. This means the context switching between active and dormant context can be achieved within nanoseconds, especially for the RESTORE operation, which can be done at sub-nanosecond range. Moreover, this context switching between SRAM and DRAM is accomplished internally without going through any external data bus, therefore saving time, bandwidth and energy. Scalability with technology means the switching time will be further reduced as the feature size

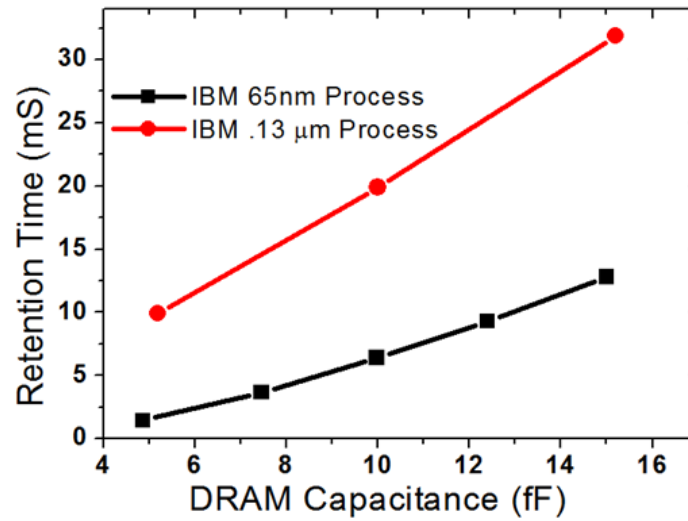


Figure 3.3 SNM and SINM for SRAMs composed of RDF, WF and Shifted WF devices.

becomes smaller. With reasonable DRAM design, the context switching can be easily achieved within one nanosecond for the current and future technologies.

3.4.3 Retention

Since the hybrid cell requires periodically refreshing the dormant context stored in the DRAM nodes, it is necessary to analyze the retention characteristics. Retention time is defined as the maximum time after the STORE operation that a successful RESTORE operation can still be achieved, which is strongly related to the EN pass transistor leakage, Q/Qb node and DRAM capacitances. The relationship between retention time and DRAM capacitance is shown in Fig. 3.3. 10 millisecond

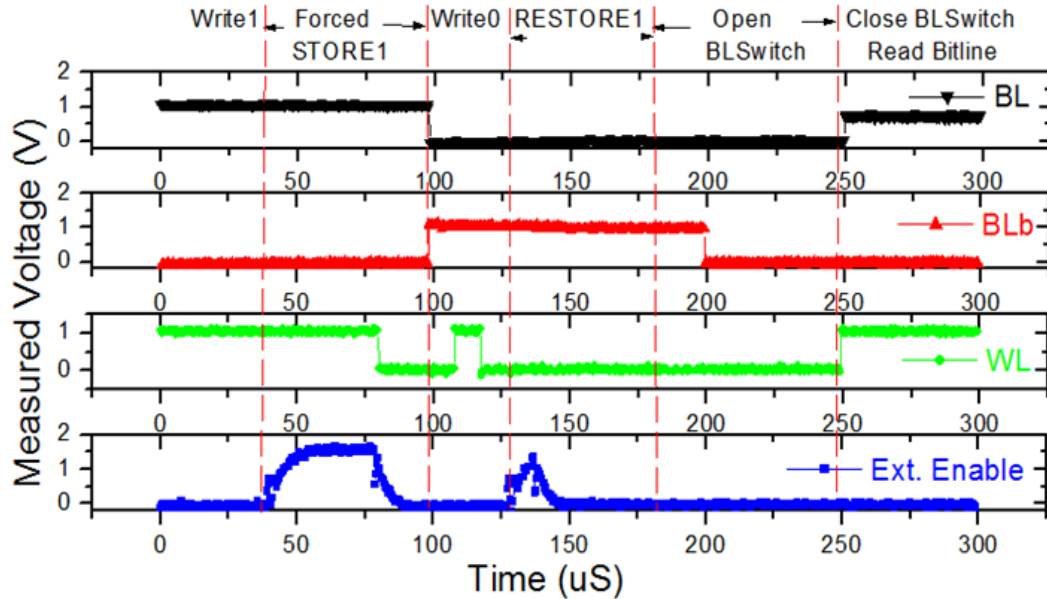


Figure 3.4 RESTORE operation testing sequence and oscilloscope measurements of BL, BLb, WL and the external enable signal. DRAM capacitors are directly charged by BL, and the process is called forced STORE. When WL is turned on at the last stage, BL resumes “1” and BLb resumes “0”, indicating a successful RESTORE operation.

retention time can be achieved with only 5fF at 130nm, and over 1 millisecond retention time is achieved at 65nm.

3.5 *Prototype Measurements*

Prototype cells were fabricated in IBM .13 μ m process with two DRAM branches and 5fF load capacitance. The EN pass transistor is the same as the SRAM pass transistor in order to reduce leakage. On-chip double-stage buffers sharpen the rising edge of the external EN signal to the nanosecond range. EN rising time is then fine-tuned through buffer control voltages. STORE and RESTORE operations were characterized individually with the translated external EN signals. RESTORE was first tested, where the DRAM capacitor was charged through bit line instead of SRAM,

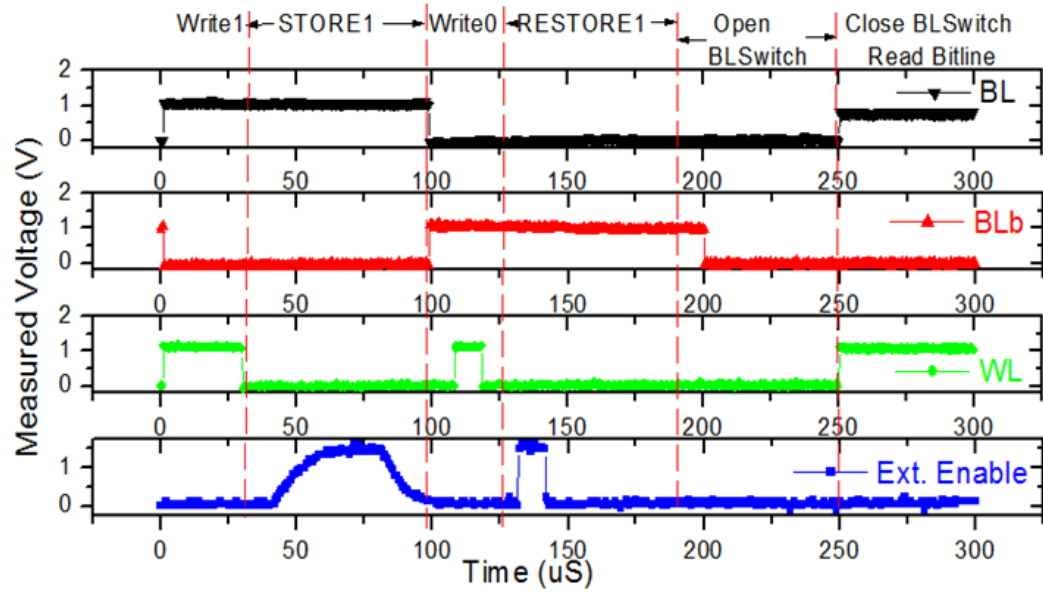


Figure 3.5 STORE operation testing sequence and oscilloscope measurements of BL, BLb, WL and the external enable signal before double buffer sharpening. RESTORE rising time is set to be fast enough to guarantee a successful RESTORE. When WL is turned on at the last stage, BL resumes “1” and BLb resumes “0”, indicating a successful STORE operation.

forcing a guaranteed context STORE operation to the DRAMs. STORE was then characterized while RESTORE is performed with confidence. Measurement sequences and oscilloscope readings are presented in Fig. 3.4 and 3.5 for the two characterizations, respectively.

Fig. 3.6 illustrates the retention measurements of at least 10ms at 5fF cross capacitance, which matches the simulated results. Critical EN rising time and retention characteristics were then gathered. The range of EN rising time that allows a successful RESTORE in the prototype cell is from 0.5ns to 4 ns as shown in Fig. 3.7, and lies within the range of the simulated results. The lower limit of 0.5ns is due to the limitation of measuring equipment rather than the cell itself. Both conventional DRAM (pairs of 1T-1C DRAM) and the 2T-1C cross capacitor DRAM configurations

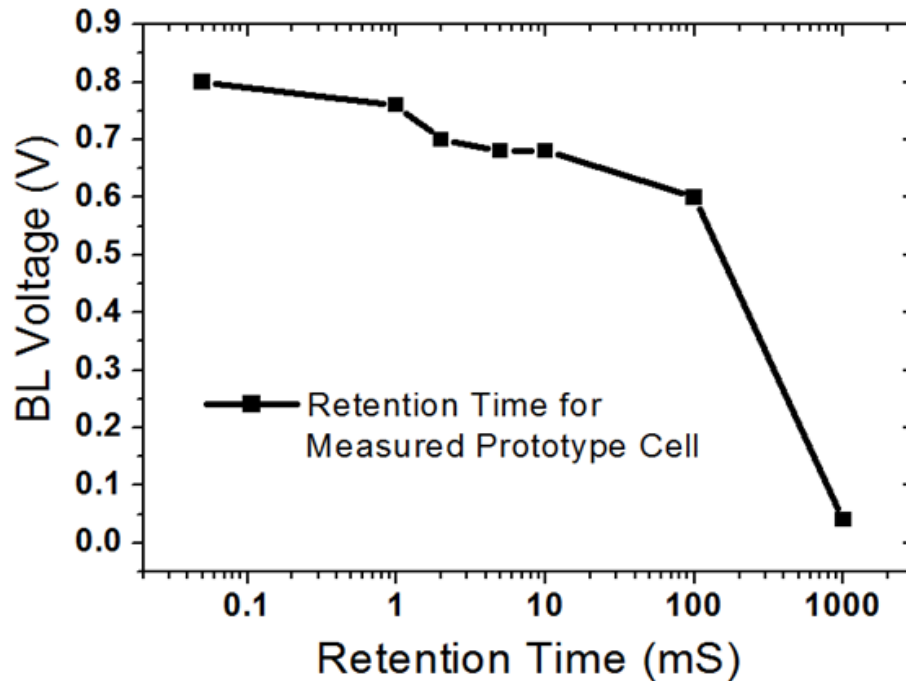


Figure 3.6 BL voltages at various retention times. 10ms of retention time can be achieved with less than 10% drop from the highest obtainable BL voltage.

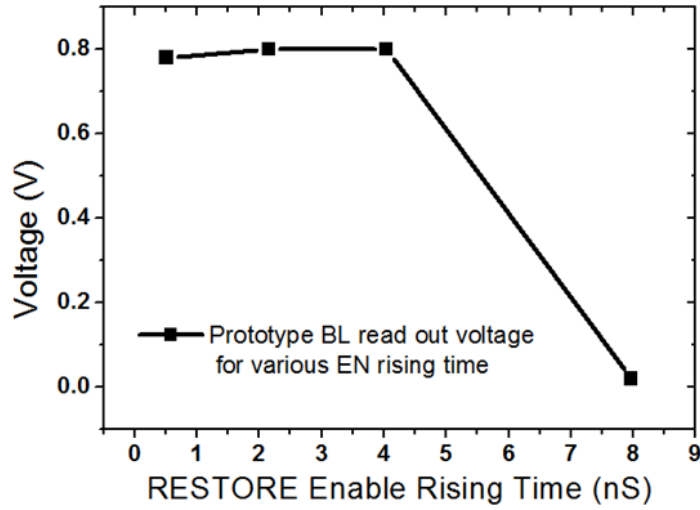


Figure 3.7 Range of enable rising time for achieving a successful RESTORE operation is around 0.5ns to 4ns. The lower bond is limited by the instrument parasitic.

were fabricated and tested. Hybrid cell with the 2T-1C cross capacitor scheme shows better tolerance to mismatches and can achieve successful STORE with EN rising time between 8ns to 100ns. On the other hand, cells using pairs of 1T-1C DRAM [4] suffered from additional mismatches, and only allowed STORE operation to be performed successfully between 8ns to 10ns, as seen in Fig. 3.8, which is significant reduction in noise margin compared to the case of the cross capacitor DRAM. This improvement in cell stability is caused by the inherent complimentary nature of the 2T-1C DRAM branch, which is more resilient to noise disturbance and shows promise in using cross capacitor for SRAM disturb stabilization.

3.6 *Disturb Stabilization*

3.6.1 *Disturb Stabilization Analysis*

When both enable transistors are shorted or remain ON, the SRAM data is

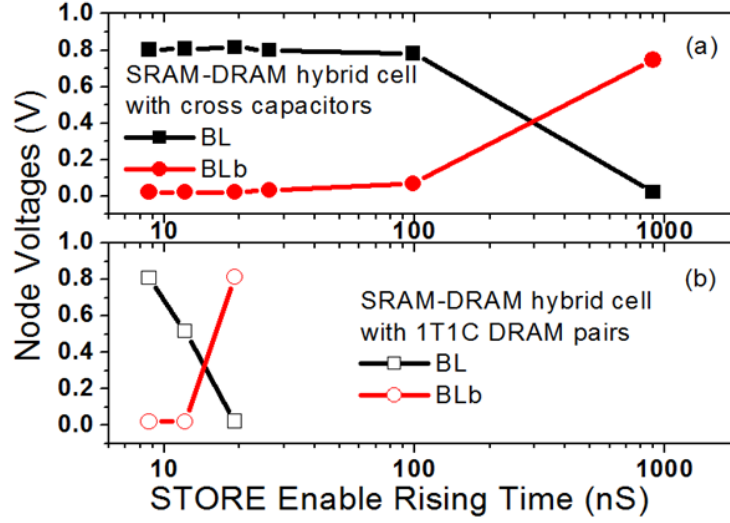


Figure 3.8 EN rising time range for the hybrid cell with (a) 5fF cross capacitance that can achieve a successful STORE operation is between 8ns and 100 ns in prototype measurements, but only 8 to 10ns when using (b) a differential 1T-1C DRAM pair due to additional mismatches. STORE does not fail in simulation if no mismatch is present.

reinforced by the differential DRAM. This is an added bonus for the hybrid cell, and can also be deliberately introduced to the conventional SRAM, depending on the appropriate applications.

This stabilization effect were first evaluated through read current noise margins (SINM) [5], where the current noise was introduced directly into the internal storage node with a linear voltage source ramped from 0 to 1V in 10ns, which is comparable to today's SRAM access time. This N-curve study is the conventional metric used to evaluate the SRAM read stability due to its simple interpretations and broad applicability. To analyze how much improvement the additional cross capacitors can contribute to the SRAM noise margin, various transistor threshold voltage mismatches

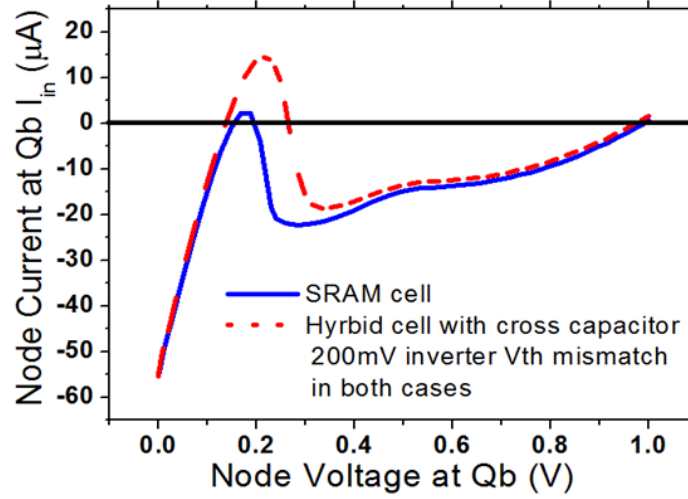


Figure 3.9 Read current noise margins (SINM) at 200mV inverter mismatches for both regular SRAM and the SRAM-DRAM hybrid cell with 5fF cross capacitor.

were introduced to the inverter pair, following the worst case scenario with stronger PFETs and weaker NFETs in different branches. Fig. 3.9 shows an extreme case where 200mV mismatch nearly flips the regular SRAM states, while one 5fF cross capacitor increases the SINM of the hybrid cell by 11 μ A. Because no intentional mismatch was originally present in the fabricated cells, prototype N-curve measurement was performed by lowering the supply voltage to 0.8V and floating the bit line in order to further exacerbate the inverter mismatch and cell instability. Due to the impracticability of measuring the Qb node current directly, the current is measured from BLb, which results in an appreciable difference in the N-curve shapes compared to conventional ones. Word line was slightly boosted in order to observe the signature three zero crossings that signify the stable states of SRAM [5]. Results are plotted in Fig. 3.10, where one 5fF load capacitor increases the SINM by about 8 μ A, which is around 20% improvement in the measured current noise margin.

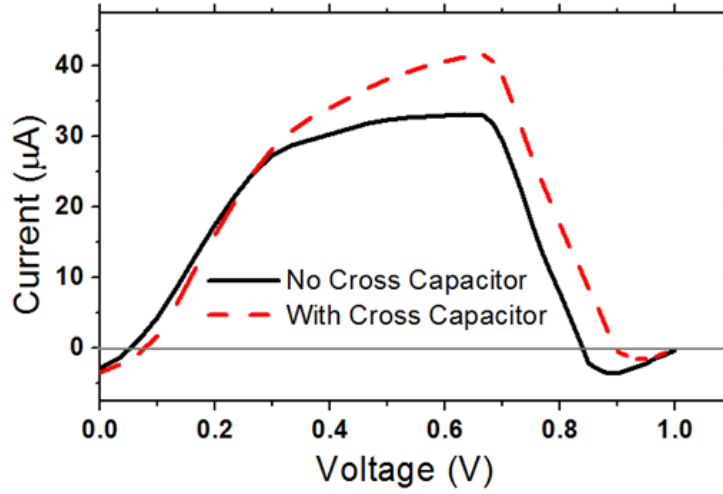


Figure 3.10 Conventional read current noise margin measurements for the hybrid cell. BLb currents were shown for EN=0V (no cross capacitor) and EN=1.5V (with cross capacitor)

However, the phenomenon of current switching direction, which should have indicated the change of the SRAM states, was not observed in some prototype cells. It also has a strong dependency on WL voltage and noise duration. This is because when DRAM cross capacitor existed between Q and Qb, extra transient leakage path is created. Fig. 3.11 illustrates the simulated transient SRAM states, BLb current, and the capacitor current obtained when DRAM EN pass transistors are OFF. Fig. 3.12 shows the corresponding results when the cross capacitor branch is turned ON, where the Q and Qb still flip even though there is no negative BLb current observed due to the additional cross capacitor current. Therefore, the zero crossings of the conventional N-curve measurement are no longer accurate representations of the internal storage states. Although the N-curve still shows the correct trend, it cannot provide appropriate predictions for the read current noise margin under transient conditions with capacitive load stabilization. Thus, it is crucial to find a more suitable noise margin metric to

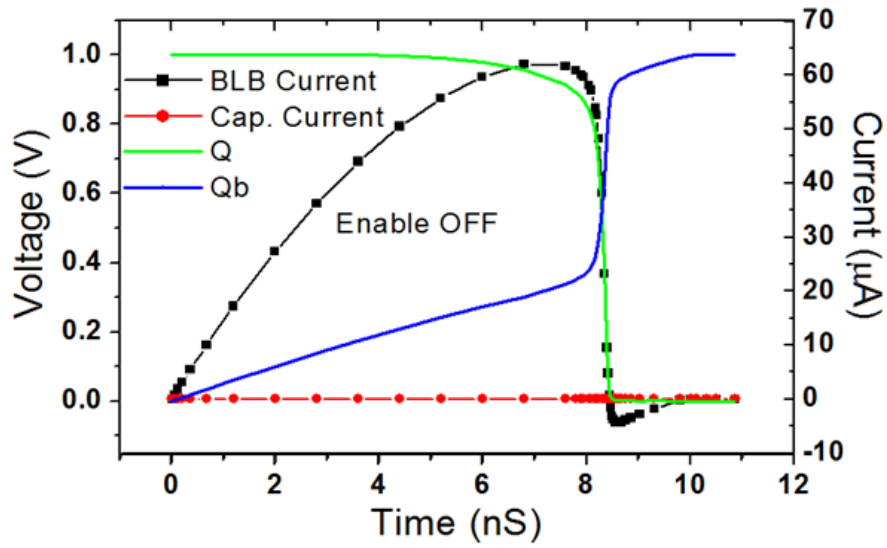


Figure 3.11 The N-curve extracted through BLb, its corresponding SRAM states and leakage current through cross capacitor with $EN=0V$.

analyze the hybrid cell stability performance in dynamic operations.

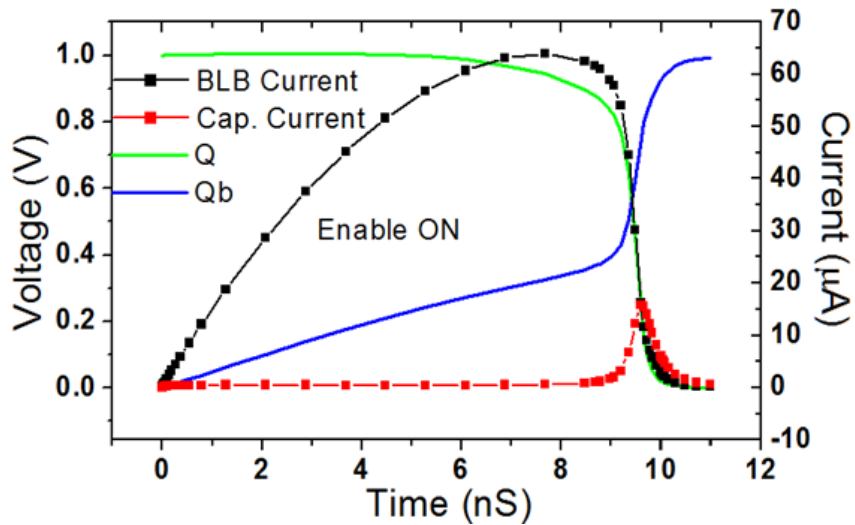


Figure 3.12 The N-curve extracted through BLb, its corresponding SRAM states and leakage current through cross capacitor with $EN=1.5V$.

3.6.2 Realization of Both Multi-bit Storage and Disturb Stabilization

Realization of both multi-bit storage and disturb stabilization may be performed with minimum cost for suitable applications. For instance, in multi-context register files [4], where there is always a DRAM branch storing the same context as SRAM, there is no need to specifically assign extra branch for stabilization. By turning on the dormant branch that has the same information of the active context during SRAM read, the only overhead will be the time spent turning on the desired DRAM.

3.7 Transient Noise Margin

Limitation of using the conventional N-curve-based noise margin metrics has

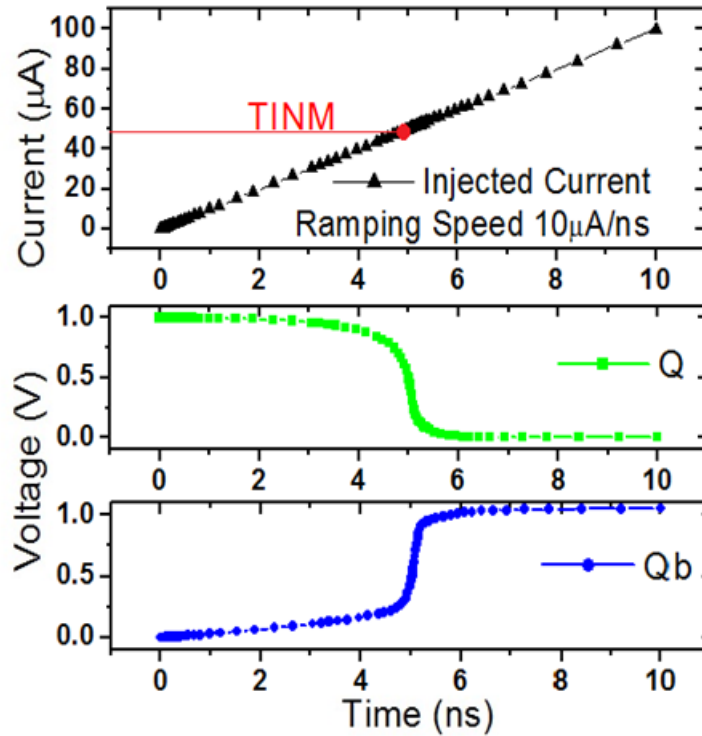


Figure 3.13 A sample illustration of the linear current source at a ramping speed of 10 μA/ns. The transient current noise margin (TINM) is defined when Q and Qb states flip.

become a major issue as SRAM scales down drastically [6-8]. Read and write operations are performed in an increasingly dynamic fashion due to the shrinking access time as well as deployment of read/write assist circuits [8]. This time-dependent aspect of SRAM operation prompts the use of realistic transient disturb for characterizing SRAM performance [6-8], because precise timing control plays a critical role in evaluating successful read and write operations. The transient read current noise margin (TINM), which is defined as the maximum tolerable current level injected through bit line before SRAM fails to maintain its state, is simulated and compared to the conventional SINM obtained from the N-curve technique on the effectiveness of cross capacitor stabilization.

A 0 to 100 μA linearly varying current source with various ramping speeds is introduced to the hybrid cell to represent time-dependent noise sources during dynamic read. This current source will not only aid in capturing the transient noise behavior when read operation is initiated, but also the difference of the injected noise

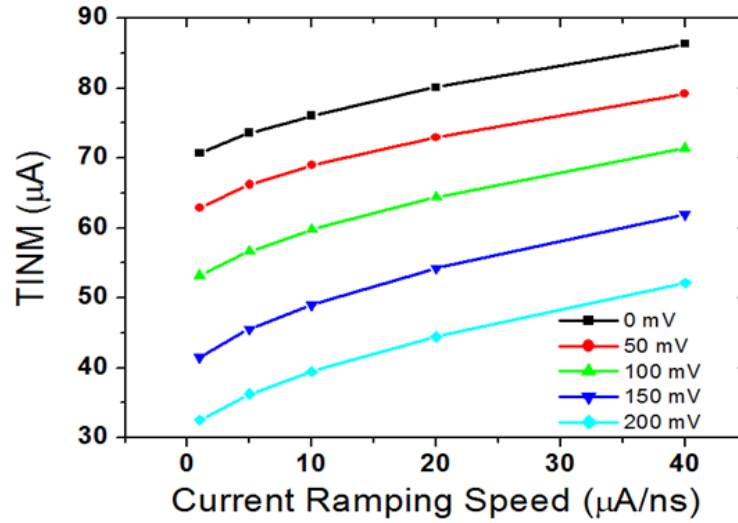


Figure 3.14 Transient read current noise margin (TINM) under various inverter mismatches and noise injection speeds with no cross capacitor ($E_N=0\text{V}$).

charges from various pulse durations, which represent scaling effect of the access time. The current ramping speed is directly related to the noise pulse duration. A slower ramping speed allows the same amplitude of noise to be injected into SRAM with a longer duration, therefore allowing more time for the internal feedback to counter the noise.

For example, $10\mu\text{A}/\text{ns}$ ramping speed means the SRAM will see a total of 0.5pC of charge injection in 10ns , which is referenced to a typical SRAM access time. On the other hand, $40\mu\text{A}/\text{ns}$ ramping speed denotes a scenario where the noise level reaches $100\mu\text{A}$ with only $\frac{1}{4}$ of the total charge injected compared to that of the $10\mu\text{A}/\text{ns}$. The TINM is then recorded at the noise amplitude when the SRAM fails to maintain its original state, as illustrated in Fig. 3.13. TINMs for various current injection speeds against inverter mismatches up to 200mV are summarized in Fig. 3.14 for a conventional SRAM without the cross capacitor. The higher the ramping

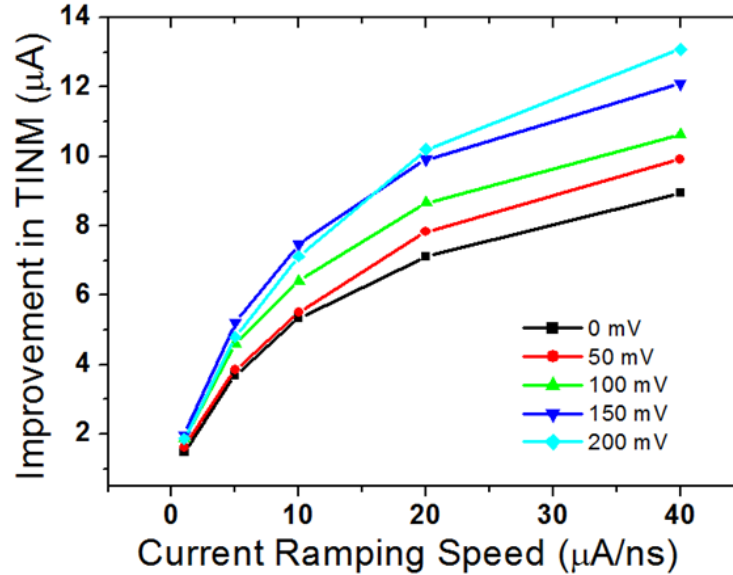


Figure 3.15 TINM differences with and without 5fF cross capacitor under various inverter mismatches and noise injection speeds.

speed, the shorter the noise pulse, and therefore the larger the read current noise margin, as it has fewer noise charges to flip the original SRAM states.

Fig. 3.15 shows the TINM improvement comparison between SRAM with and without the 5fF cross capacitor. Improvement on the TINM from the cross capacitor scales with the noise injection speed, meaning the enhancement in SRAM read stabilization actually increases as the access time scales down. TINM also improves when the cross capacitance increases due to more charges stored on the hybrid cell, making it more difficult to disturb the states, as indicated in Fig. 3.16. Both SINM and TINM metrics show similar trend, with SINM smaller than TINM, as illustrated in Fig. 3.17.

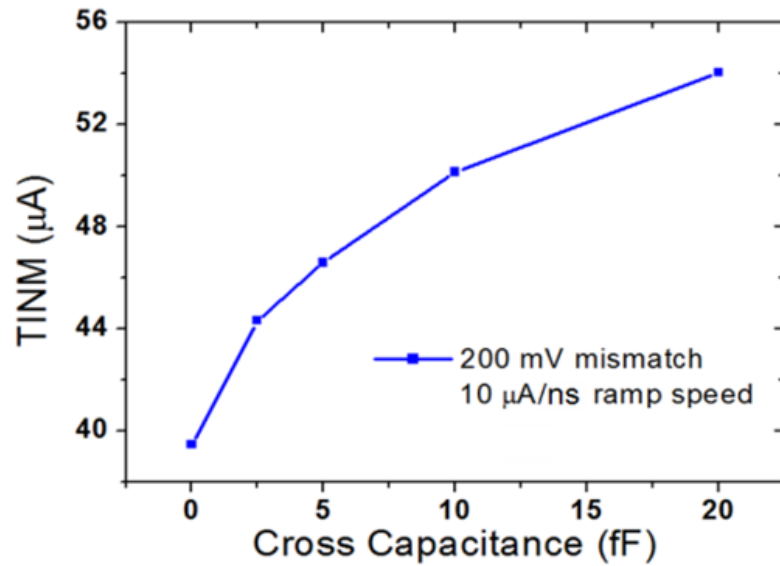


Figure 3.16 TINM for various cross capacitances with 200 mV inverter pair mismatch and 10ns noise duration.

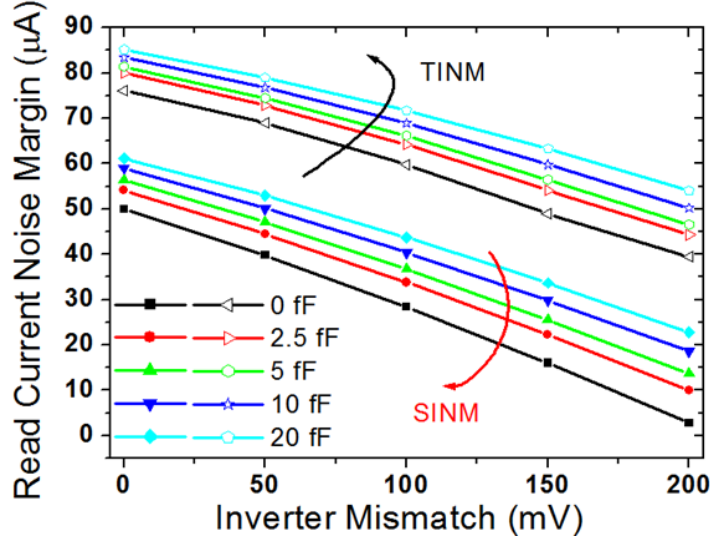


Figure 3.17 SINM and TINM with various inverter mismatches and capacitances for 10ns noise duration.

3.8 Conclusion

In this brief, circuit characteristics of the hybrid SRAM-DRAM cell with cross capacitors are analyzed. The integrated cell is capable of storing multiple contexts within one cell and achieving nanoseconds internal swapping between active data in SRAM and dormant data in DRAM nodes. Data movement direction depends on the critical EN signal ramping time, which scales with technologies and cross capacitances. At 5fF load capacitance, we achieved 10's ms of retention time in the DRAMs, small increase in access delay, and energy and space saving for context-switching applications.

Prototype cells were fabricated and the measurements agree well with simulation predictions. Immunity to read disturb and mismatches was observed in the

hybrid cell with cross capacitors. Improvement in read current noise margin is evident in both simulation and measurements. Limitation of the conventional N-curve SINM is analyzed and compared to TINM.

REFERENCES

- [1] F. Hamzaoglu, K. Zhang, Y. Wang, H. J. Ahn, U. Bhattacharya, Z. Chen, Y-G. Ng, A. Pavlov, K. Smits and M. Bohr, "A 153Mb-SRAM design with dynamic stability enhancement and leakage reduction in 45nm high- κ metal-gate CMOS technology", *ISSCC Dig.*, pp. 376-377, Feb. 2008.
- [2] K. Chun, P. Jain, T-H. Kim and C. H. Kim, "A 667 MHz logic-compatible embedded DRAM featuring an asymmetric 2T gain cell for high speed on-die caches", *IEEE J. Solid-State Circuits*, vol. 47, no. 2, pp. 547-559, 2012.
- [3] A. Valero, J. Sahuquillo, S. Petit, V. Lorente, R. Canal, P. L'opez and J. Duato, "A hybrid eDRAM/SRAM macrocell to implement first-level data caches", *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, pp. 213–221, 2009.
- [4] W. Yu, R. Huang, S. Xu, S.-E. Wang, E. Kan, and G. E. Suh, "SRAM-DRAM hybrid memory with applications to efficient register files in fine-grained multi-threading", *Proceedings of the 38th Annual International Symposium on Computer Architecture*, pp. 247-258, 2011.
- [5] E. Grossar, M. Stucchi, K. Maex and W. Dehaene, "Read stability and write-ability analysis of SRAM cells for nanometer technologies", *IEEE J. Solid-State Circuits*, vol. 41,no. 11 ,pp. 2577-2588,Nov. 2006.
- [6] J. Wang, S. Nalam and B.H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs", *ISLPED '08*, pp. 129-134, Aug. 2008.
- [7] S. O. Toh, Z. Guo and B. Nikolic, "Dynamic SRAM stability characterization in 45 nm CMOS", *Symp. VLSI Circuits Dig.*, pp. 35 -36, 2010.

- [8] W. Dong, P. Li and G. M. Huang, "SRAM dynamic stability: Theory, variability and analysis", *ACM International Conference on Computer-Aided Design*, pp. 378-385, Nov. 2008.

CHAPTER 4

IMPLICATIONS OF VARIATION SOURCES ON FLASH PHYSICAL UNCLONABLE FUNCTION DESIGN AND USAGE

4.1 *Abstract*

Universal process variations in Flash physical systems are identified and decomposed into layout, intrinsic, stress and bit-wise fluctuation sources. The study shows the understanding of systematic variations and noise sources are essential for improving security and reliability of FPUFs. Bit-wise variations are proven to be originated mainly from random dopant fluctuation, which is indeed truly random and impossible to duplicate. Overall, this chapter provides a theoretical foundation of FPUFs whereas previous PUF studies rely only on experimental evidence for its security and entropy. Algorithms to generate crypto keys from Flash Physical Unclonable Functions (FPUFs) for security applications are illustrated. Corresponding maximum entropies are calculated. The fact that FPUF can generate as many crypto bits as plaintext without much penalty shows promises in a more secured data coding scheme. A simple example is used to demonstrate the flexibility of manipulating FPUF bits to quickly generate large amount of crypto keys, although a more advanced algorithm is required for actual security applications, which will leave to become future work.

4.2 *Introduction to Physical Unclonable Functions (PUFs)*

Researchers have recently demonstrated that inherent manufacturing variations can be exploited to authenticate an IC instance or generate unique secrets for each chip. This primitive is named Physical Unclonable Functions (PUFs). PUF is a physical one-way function that provides unique challenge-response pairs based on the intrinsic,

uncontrollable but reproducible randomness of the implementing device. First set of PUF implementations are based on optical interference patterns [14]. Later on, Silicon based PUFs have gradually gained popularity due to its CMOS compatibility and more general applications in electronic systems.

PUF implementations are traditionally classified into two general types. First is delay-based PUFs that utilizes the different digital circuit delay of seemingly identical pathways, such as ring oscillator [4, 19] and arbiter [10] PUFs. Second category is memory-based PUFs that based on the mismatched bistable cells whose behavior is affected by the intrinsic process variations. This type of PUFs include SRAM [5, 6], latch [9, 18], D flip-flop [11] and buskeeper [17] PUFs. Recently, a different type of memory PUF based on Flash memory (FPUF) has been proposed [15, 22, 23], which exploits the program/erase time variations among Flash cells.

So far, studies on PUFs have been largely focused on experiments to demonstrate that there exist enough variations to distinguish individual chip fingerprints. Few detailed characterizations of the physical mechanisms behind the variations have been incorporated into these studies [6, 22]. Unfortunately, without concrete definition and modeling of the physical variation sources, it is almost impossible to generalize the experimental conclusions to a variety of different devices, circuits and systems. In particular, technologies used to implement silicon based PUFs change very quickly due to the drastic scaling of the feature size. Only by physical modeling of the PUF physical origins, we can guarantee that proper PUF characteristics can be extended to different technology nodes, and other manufacturers.

This chapter presents a semiconductor device level modeling and analysis to understand underlying physical mechanisms behind variations in FPUFs, and discusses their implications for designing secure and reliable protocols. To the best of our knowledge, this study is the first to underpin the physical mechanisms of FPUFs.

Two major categories of variations are discussed: manufacturing and operational. Manufacturing variations are important in determining the uniqueness and consistency of FPUF. This type of variation sources first include layout and design-induced systematic fluctuations that are highly correlated among similar chips, reducing the overall independency of the FPUF bits. The other intrinsic random variations mainly originate from random discrete dopant fluctuations that are impossible to duplicate in current technologies. Variations caused by the field operations present main challenges for making FPUF time invariant, but enable steganography by introducing user-defined biases [23]. Major variation sources include cyclic endurance aging effects caused by program/erase stress, bit-wise fluctuations caused by random telegraph noise (RTN), and block-level erase effects. All of these variations can significantly impact the proper use of FPUF, and discussions are given to improve the system reliability, consistency and security.

The rest of the chapter is organized as follows. Section 4.3 provides an overview of the FPUF basics and comparison with other conventional PUF implementations. Section 4.4 discusses about manufacturing variations mentioned in Chapter 1 and implications on FPUF designs. Section 4.5 illustrates the variations in the field and methods to reduce FPUF fluctuations and improve consistency. Section 4.6 illustrates the quantization algorithms and the corresponding bit entropies. Section 4.7 discusses a simple example of utilizing FPUF in data encryption. Finally section 4.8 concludes the paper.

4.3 Overview on Flash Physical Unclonable Function (FPUF)

The concept of FPUF is first suggested in [15] for producing signatures to authenticate chips and generate cryptographic keys. It is especially attractive because, unlike prior PUFs, FPUFs do not require any custom hardware circuits, and the Open

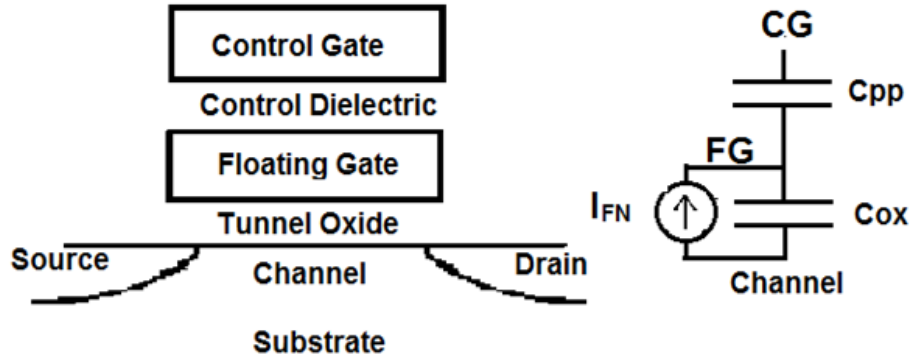


Figure 4.1 Flash memory cell schematics and equivalent operating circuit.

NAND Flash interface (ONFi) [13] sufficiently provides universal extraction methods for most of the commercial Flash chips. Experiments on producing FPUFs have been successfully carried out on commercial off-the-shelf (COTS) components [22].

FPUFs can be extracted based on a technique called partial or aborted programming. Figure 4.1 illustrates the schematic of a conventional NAND Flash memory cell and its equivalent circuit. During programming operation, Fowler-Nordheim tunneling current (I_{FN}) injects charges on the floating gate, increasing the threshold voltage (V_{th}) of the underlying transistor. When enough charges are stored in the floating gate, V_{th} of the specific cell will change to produce different levels of channel current, thus differentiating the stored state. Larger channel current is defined as a binary “1”, while smaller channel current as binary “0”.

The initial V_{th} without charge stored on the floating gate is different from cell to cell due to process variations. V_{th} difference also induces changes in I_{FN} and hence the specific program time. Hence, each individual cell will require different number of partial program pulses to change state. This partial program number varies distinctively from bit to bit, page to page and chip to chip [11], and therefore can be

Table 4.1 Bit capacity per area comparisons for various PUFs

PUF Type	Composition	PUF Bits per mm ²
NAND Flash	4GBits	3.3×10^7
SRAM	262144 SRAM cells	1230723
Buskeeper	16384 buskeepers	215579
Latch	32768 latches	120470
D Flip-Flop	32768 flip-flops	83592
Ring oscillator	4096 inverter chains	15934

considered as a unique PUF function.

FPUFs have several practical advantages over other PUF implementations. In addition to the wide applicability, FPUF extractions do not require a power cycle compared to the PUFs based on bi-stable elements [5, 6, 9, 11, 18]. Since Flash is one of the most aggressively scaled technologies, FPUF is also superior in bit capacity compared to other PUFs. A sample comparison is presented in Table 1 for 65nm technologies; results are partially calculated from [8].

4.4 *Manufacturing Variations in Flash PUF*

4.4.1 *Layout Variations*

PUFs need to generate outputs that are unique and reproducible for each chip. Ideally, each PUF bit should be independent and random. Thus, any systematic

Table 4.2 Flash chips tested for FPUF investigations

Manufacturer	Part Number	Capacity	Technology
Hynix	HY27US08281A	128Mb	90nm SLC
Hynix	HY27UF084G2B	4G	50nm SLC
Micron	MT29F2G08ABA EAWP-IT:E 4	2G	34nm SLC

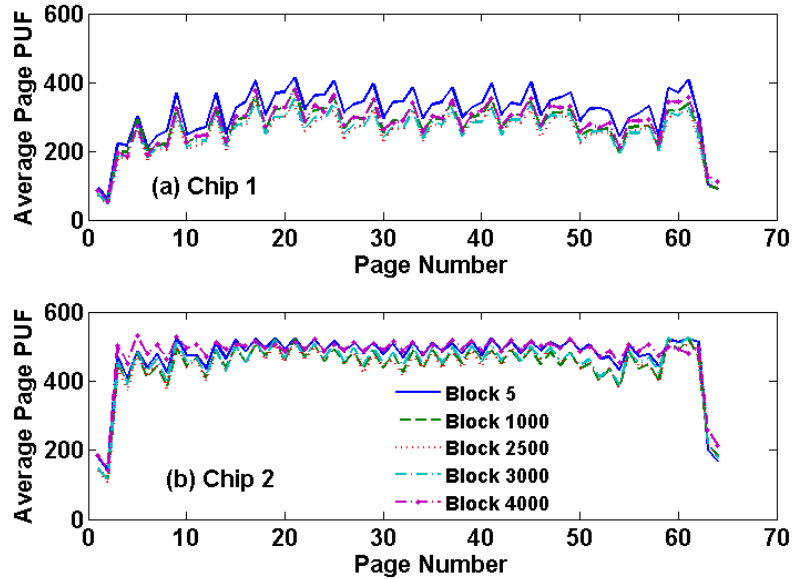


Figure 4.2 Systematic layout variations in two Flash chips with the same part number. Average page PUF is produced for each of the 64 pages for (a) chip 1 and (b) chip 2. The similar page to page fluctuation with high correlation is observed, even when the chips are not from the same wafer or lot.

variations presented in the PUF bits should be carefully examined and extracted. The source of the randomness of the FPUF should also be identified for better understanding of the technology independency as well as technology scaling effects.

Manufacturing variations will be first discussed, which refer to the variation sources that originate from the initial fabrication process. Variations caused by field variations will then be presented.

The experiments in this study are mostly performed on Hynix SLC chips due to the available technology nodes for technology scaling analysis. Micron SLC chips are used for comparison of results from a different manufacturer.

For each Flash chip, the average page FPUF is obtained by averaging the bit-level partial program pulse numbers for a particular page. This average page PUF is

then plotted for every page across the same block for several blocks. Results from two identical Hynix 50nm chips are presented in Fig. 4.2.

A consistent systematic variation can be seen among average page FPUFs for all the blocks. If a specific page in block 5 requires higher-than-average partial program number, the corresponding page in all other block also requires higher average partial program numbers. This cyclic fluctuation is highly correlated among blocks in the same chip with a correlation coefficient as high as 0.99. Similar fluctuation patterns have been found in other chips with the same part number, and the correlation coefficient is around 0.88.

Because this page-wise fluctuation is consistent with all the blocks and also highly correlated for various chips, the contributor of this systematic variation has to come from a universal variation source that is the same for all blocks and chips fabricated in the same manufacturing process. Since it cannot come from spatial variation alone that varies from chip to chip, it is reasonable to attribute this prominent systematic effect to the layout design that are unique to the same group of products.

4.4.2 *Spatial Variations*

Besides layout variations, spatial variations caused by topography interaction with the fabrication process can contribute to extra biasing on the FPUF distributions.

Spatial variations come from sources such as layer deposition gradient and pattern planarization in chemical mechanical polishing [3], which can result in intra-die variations. These chip-level deviations highly depend on the die locations on their corresponding wafers, and can cause substantial FPUF biasing from chip to chip.

In order to observe the spatial variation components of the FPUFs, block averages across three similar chips are plotted in Fig. 4.3. Each curve represents the average partial program pulse number for over 4000 blocks throughout the chip.

Similar parabolic shapes and small fluctuations in similar block addresses are present for all chips.

The fluctuation in the block average FPUF across the chip is highly correlated among all three chips, with an average correlation coefficient of 0.76. Because it is very unlikely that all three chips come from the same spatial wafer location, the high

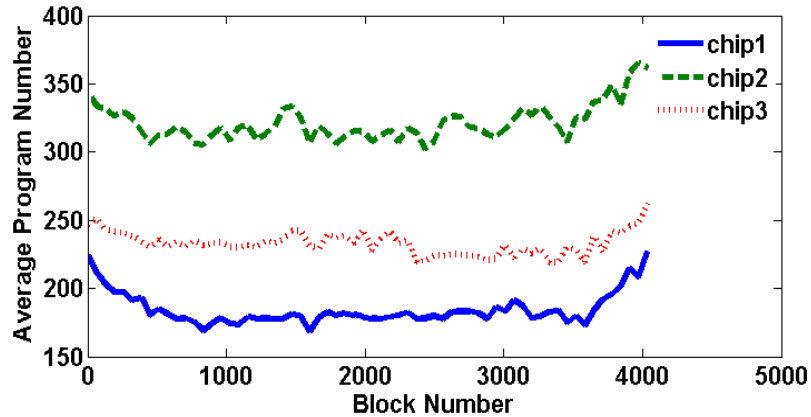


Figure 4.3 Fluctuations across 4000 blocks in three chips of the same part number but from different production lots and wafers. The three curves show high correlations in general shape and fluctuations due to layout. The large off-set in curve locations, however, is attributed to spatial variations.

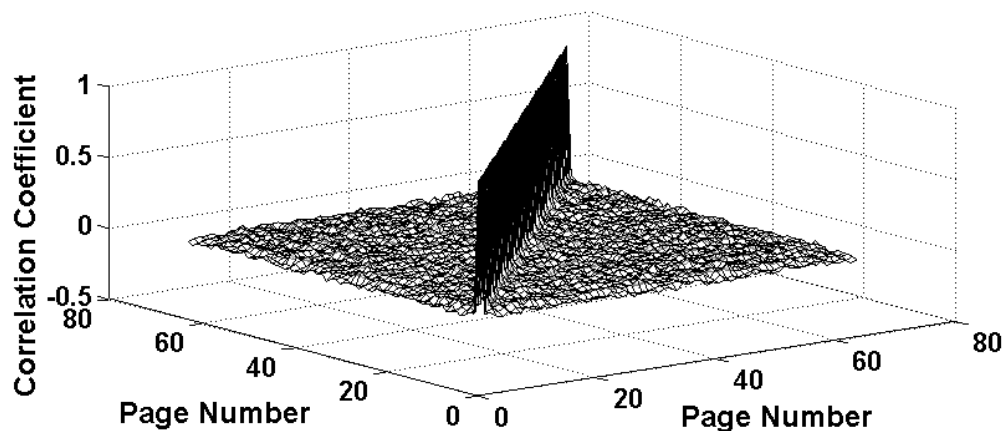


Figure 4.4 Correlations of Flash fingerprint measurements obtained from the same page and with different pages.

correlation should be additionally attributed to the systematic layout-induced variations, as previously discussed. However, there is a clear offset in the block average among three chips. This offset is not a layout systematic component and very likely comes from the die spatial location difference, which can be attributed to a spatial variation component of the manufacturing variations.

4.4.3 *Intrinsic Variations*

Flash page-level fingerprints are unique and robust enough to be used to authenticate individual chips [15, 22]. It has been reported that the average correlation coefficient for the same page is on the order of 0.97 and fingerprints extracted from different pages, either the same page from different chips or different pages from the same chip, have an average correlation coefficient around 0.0076 [22]. Fig. 4.4 illustrates an example of FPUF correlations between two measurements on the same block. Only the FPUFs extracted from the same page have correlation coefficients close to 1, and FPUF correlations between different pages are all close to 0.

Previous studies [15, 22] all simply contributed this randomness to general intrinsic process variations of Flash memory, but no detailed characterization and modeling have been performed. It is crucial to understand this source of random variation in order to determine if the uniqueness and robustness of the FPUF are indeed universally applicable, and not just a phenomenon presented in the limited selection of Flash chips.

Some of the candidates include the predominant intrinsic variation sources in sub-100nm MOS devices: random dopant fluctuation (RDF) and line edge roughness (LER), as mentioned in Chapter 1. RDF refers to the random fluctuation of the relatively small number of dopants and their discrete microscopic arrangement in the channel region, which will lead to significant variations in threshold voltage and drive

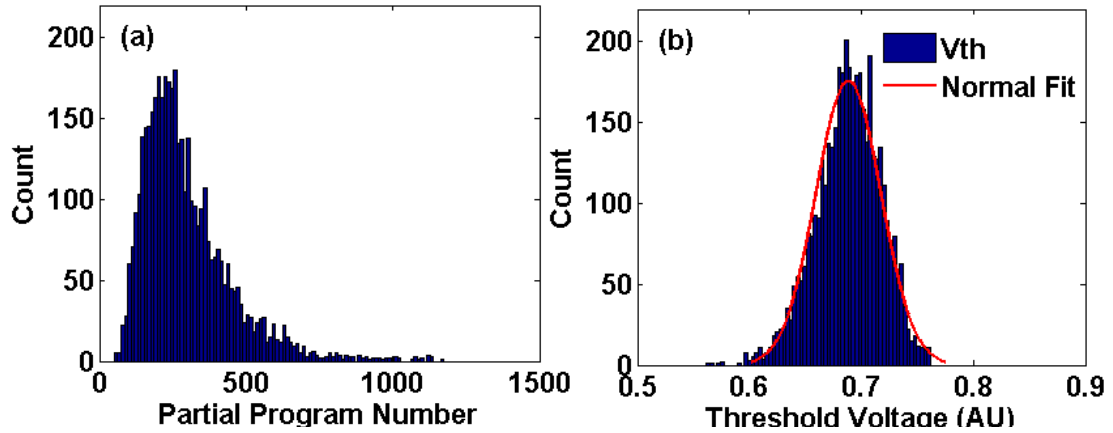


Figure 4.5 Example of (a) FPUF distribution and (b) translated threshold voltage distribution with Normal fitting.

current. [1,24,25]. The LER is caused by tolerances inherent to materials and tools used in the lithography process [25]. When considering the total intrinsic variations, RDF dominates the overall contributions before the 20nm technology node. When technology scales further than 20nm, additional variation sources are very likely originated from LER [25].

Due to the additional floating gate and control dielectric, as shown in Fig. 4.1 and mentioned in Chapter 1, Flash memory can suffer more severely from RDF compared to conventional logic devices due to the larger effective oxide thickness (EOT) from the control gate. In order to analyze the physical origins of the randomness, FPUF distributions are translated to threshold voltage distributions, and then fitted to RDF analytical models [1, 24]. Detailed analyses are listed in the appendix. Fig. 4.5 confirms the translated V_{th} distribution from FPUF resembles an RDF Gaussian distribution. However, since some other noise source can also cause data to be similar to a Gaussian shape, the scaling effect of RDF is extracted.

Three Flash chips from different technologies are characterized, which are 34nm, 50nm and 90nm. Since RDF is more severe as device scales further, a device

with a smaller feature size should result in a larger standard deviation. According to Equation 1.1, device from these three generations should increase as $1/\sqrt{WL}$ decreases. The extracted threshold voltage distributions are graphed in Fig. 4.6 with their respective normal fittings. Chips from the 34nm technology shows largest spread in the threshold voltage distributions, just as predicted. The expected standard deviations of RDF induced threshold voltage variations and the experimentally extracted ones are compared and plotted in Fig. 4.7. It is clear that the two scaling trends match very well, confirming that up to 34nm nodes, the major contributor to the overall Flash process variation is still RDF. This is because LER, which should become significant around 20nm for logic devices [25], is less severe for Flash memory due to the reduced sensitivity between gate and channel.

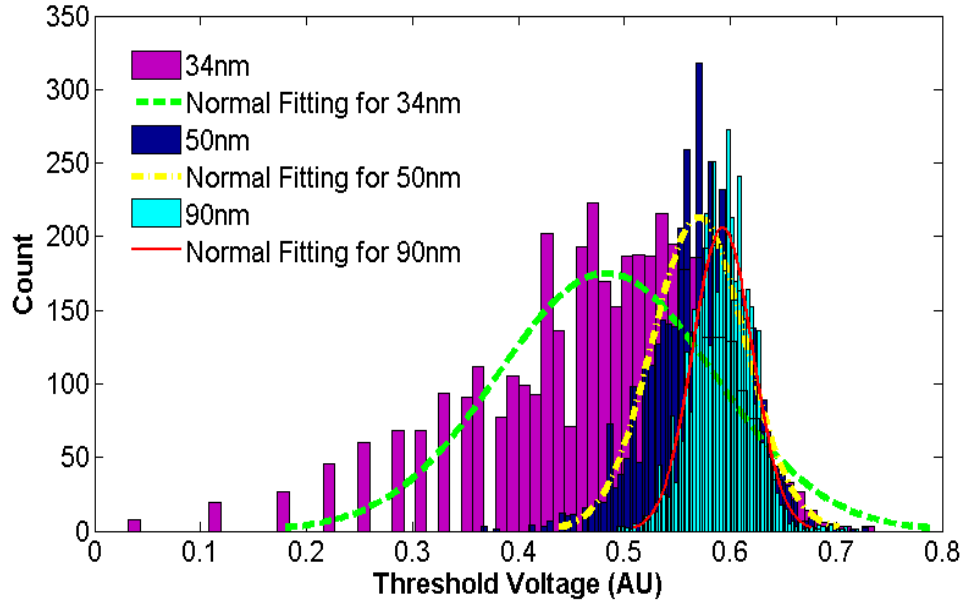


Figure 4.6 Translated threshold voltage distributions from three technology generations and the respective normal fittings.

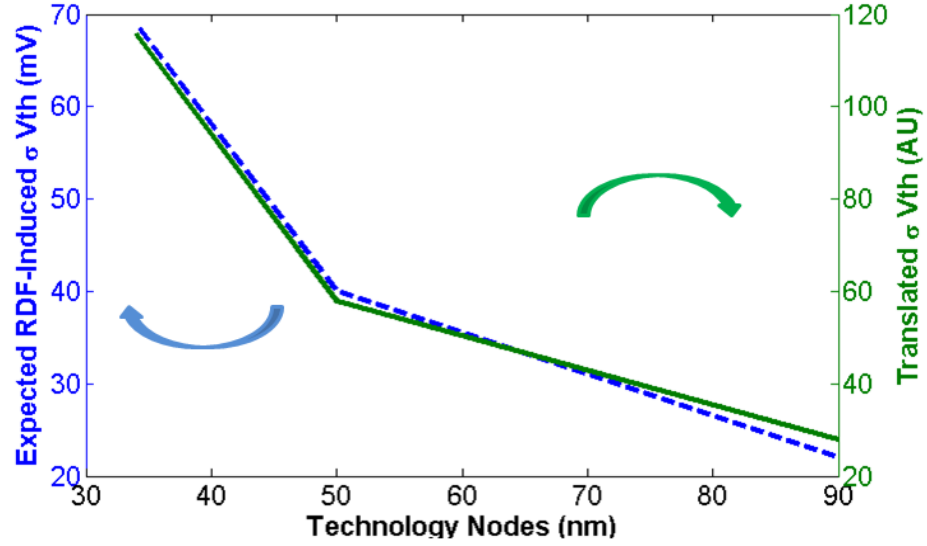


Figure 4.7 Expected scaling trend of RDF induced threshold voltage variations [25] and experimental translated ones from Flash chips measurements vs different technology nodes. The translated threshold voltage distribution is in arbitrary unit (AU) because of the additional gate coupling ratio for Flash memory that is not there for logic devices [25], which generally results in larger threshold voltage distributions caused by the lack of gate control to the channel due to additional floating gate and control dielectrics. However, the scaling trend is still valid to compare between the two CMOS devices among various technology nodes.

4.4.4 Implications in FPUF design

Systematic variations can degrade the uniqueness and entropy of FPUFs. In order to assess the effect of the layout and design induced variations, Diehard randomness tests [12] are performed for FPUFs containing systematic components and for FPUFs with systematic component removed from their page average. The tests are performed between FPUFs from the same chip and FPUFs created by concatenating

Table 4.3 Diehard tests on FPUFs with and without systematic components.

Test Type	P values for FPUFs with and without systematic variations			
	FPUFs from same chip		Concatenated from multiple chips	
	With	Without	With	Without
OQSO	1.0000	.9231	1.0000	.5414
	1.0000	.5695	1.0000	.9045
	1.0000	.9539	1.0000	.8659
DNA	1.0000	.5906	1.0000	.8440
	1.0000	.7040	1.0000	.3938
	1.0000	.2569	1.0000	.0250
	1.0000	.6446	1.0000	.0507
	1.0000	.4831	1.0000	.8076
	1.0000	.9977	1.0000	.8180

same-page FPUFs from multiple chips with highly correlated systematic variations. Sample p-values are shown in Table 4.3. OQSO (Overlapping-Quadruples-Sparse-Occupancy) and DNA are two randomness tests from the Diehard suite, and p value close to one suggesting data fail the specific category. By removing the systematic components from the FPUF bits, improvement on the randomness is evident in both cases of OQSO and DNA tests.

On the other hand, determining RDF as the major source of FPUF variations proves that FPUFs can be extracted from any Flash memory chips on bulk CMOS processes, since it is a universal phenomenon and cannot be fully controlled or duplicated by today's fabrication technology. Extensive modeling attempts have also been conducted to recreate RDF effects [17, 18], but today's computing power is still insufficient for carrying out 3-D "atomistic" simulations on a large statistical scale to accurately depict the RDF behavior. Therefore, cloning the FPUFs that originate from RDF effect will be extremely difficult.

4.5 Variations in the Field

4.5.1 Stress-induced Variation

Uniqueness and reproducibility are both important for reliable use of FPUFs in practical settings. Reproducibility accounts for how well a certain FPUF can produce the same output response for the same input challenge when noise sources are presented in the physical system [21]. This section will discuss the field variation sources for Flash memory that may reduce the reproducibility of the FPUFs, and potential solutions to improve their reliability. Previous studies [22, 23] have suggested that repetitive program and erase (P/E) cycles can alter the partial program time of Flash cells due to cyclic endurance aging effects, thus changing its device

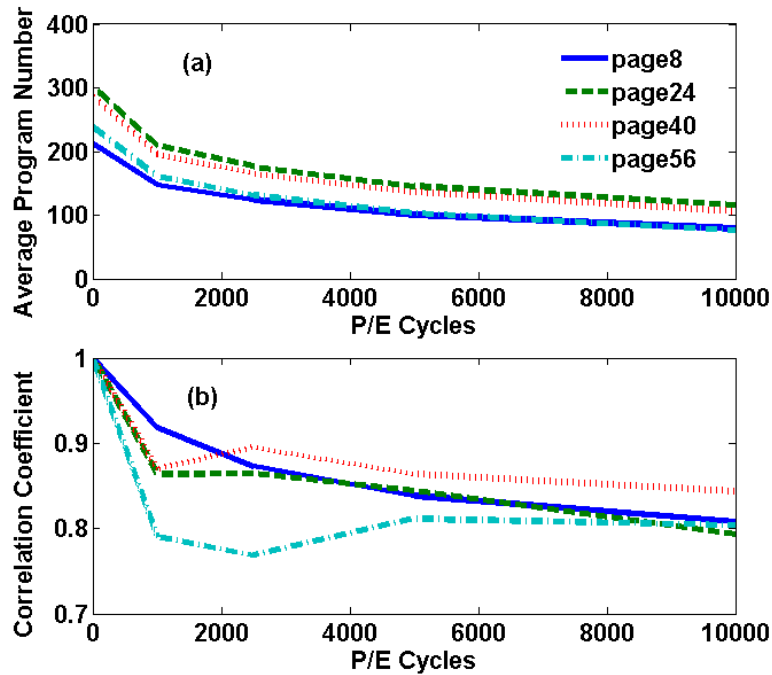


Figure 4.8 P/E stress effect on (a) average partial program numbers and (b) FPUF correlation coefficients.

signature. This stress effect can be isolated from the manufacturing variations because RDF and LER should not be directly affected by P/E stress.

Fig. 4.8 (a) illustrates how page average changes as P/E cycles increase. Correlation coefficients of fresh and stressed FPUFs from the same pages are plotted in Fig. 4.8 (b). The decrease in correlation suggests that FPUFs are becoming increasingly different as the P/E stress level rises.

Decrease of the average partial program number suggests increasing of the tunneling current, which is very likely caused by the additional trap-assisted tunneling and bias-temperature instability (BTI). Generation of trap site as P/E stress increase is consistent with the model of stress induced leakage current (SILC) [7].

4.5.2 *Random Telegraph Noise*

Although the FPUF responses are reasonably unique for different chips from correlation studies, when the same PUF bits are measured multiple times, the bit-wise partial program times have non-negligible fluctuations as depicted in Fig. 4.9. This

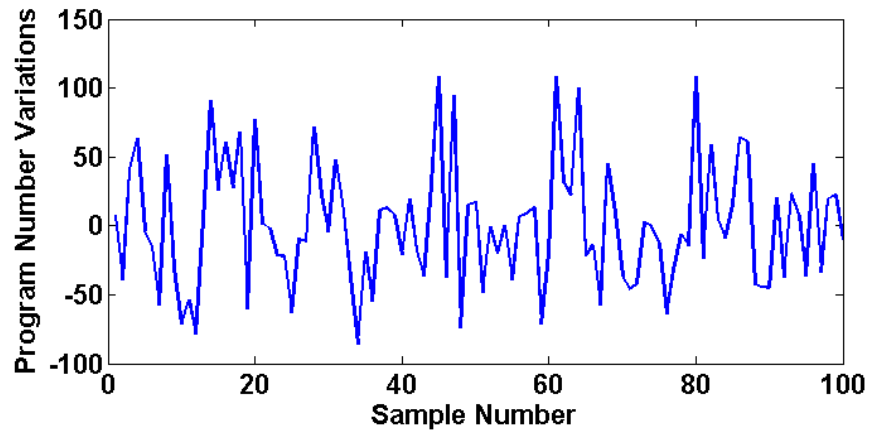


Figure 4.9 Bit-wise fluctuations for multiple FPUF measurements.

can affect the reproducibility of FPUFs when used as crypto keys or authentication.

An example of bit-wise fluctuation is analyzed via power spectral density; results are plotted in Fig. 4.10 (a). A clear $1/f^x$ relationship can be observed, and line fitting yields an x value around 1.8. These x power coefficients were then extracted for multiple bits within a page and the results are shown in Fig. 4.10 (b). The average x is around 1.7 and most of the values are within 1 to 2. This proves that bit-wise fluctuations in general display shot noise behavior. In addition, this relationship closely resembles a $1/f^2$ characteristic. This fits the profile of random telegraph noise (RTN) behavior very well, especially for low frequencies [20].

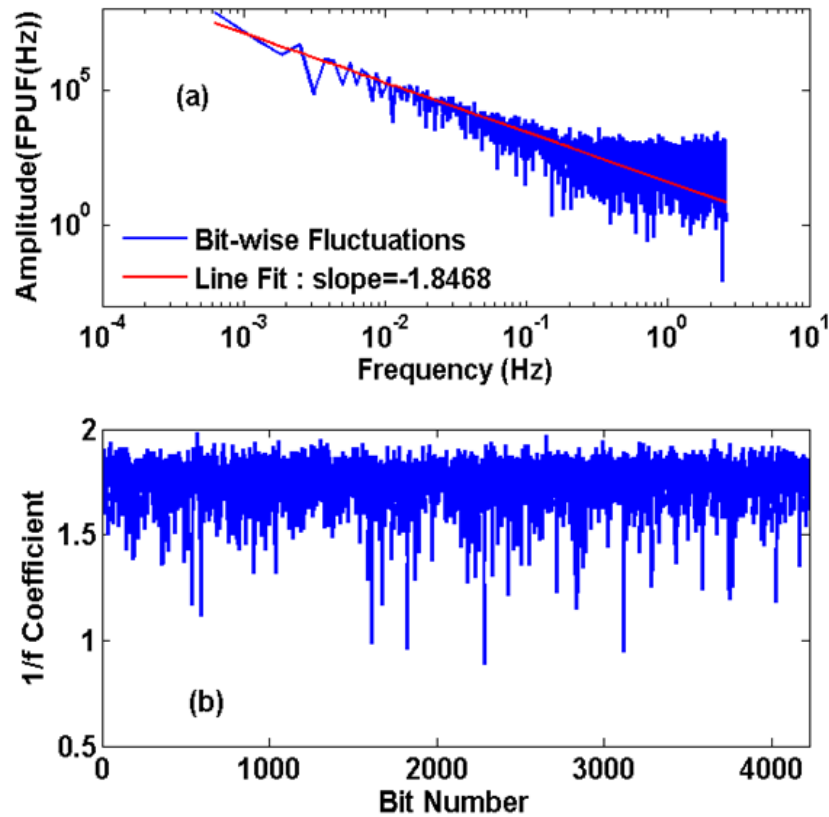


Figure 4.10 (a) Power spectral density of bit-wise fluctuations and (b) corresponding $1/f^x$ coefficients.

4.5.3 Block-level Erase Effect

We present a NC memory configuration. During the experiments carried out in section 4.2, some bits fluctuate together in the same measurement. Sample correlation coefficients on how these bits fluctuate among 5000 FPUF measurements are plotted in Fig. 4.11. A substantial percentage of bits have fluctuation correlation coefficients around 0.5.

If the bit-wise fluctuation is purely due to RTN as previously discussed, each bit should fluctuate independently. This suggests that there is additional global variation source presented. Observation from the same experiments have confirmed that conventional full erase operation dynamically adjusts the erase time during each block erase, and therefore can add a global bias to the bit-wise fluctuations. For confirmation, an erase operation with fixed erase time was performed in order to remove the block erase bias. The resulting bit-wise fluctuation correlations are also plotted in Fig. 4.11. The correlation drops significantly compared to using

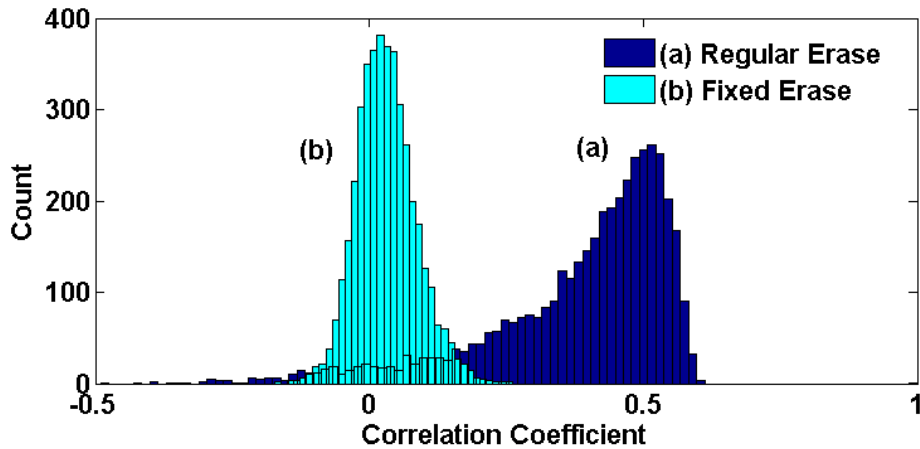


Figure 4.11 Bit-wise fluctuation correlation distributions obtained with (a) regular erase or (b) fixed erase.

dynamically adjusted full erase operations.

4.5.4 *Implications for FPUF design*

Bit-wise fluctuation originating from RTN is globally presented in all Flash memory devices and cannot be fully eliminated through current fabrication processes. Although this RTN feature is inherently useful in learning parity with noise (LPN) protocols [2], when FPUFs are directly used from a bit-wise perspective, reduction of this variation over time will be crucial to improving the reliability and reproducibility of FPUFs.

Because the origin of this variation is RTN, the variations are generally split into two levels caused by capture and emission of the trapped carriers [20]. When measured multiple times, the probabilities of FPUFs fluctuating upwards or downwards usually cancel out over time, which provides means of reducing the FPUF variations through averaging several measurements from the same physical system. This RTN effect also illustrates the improper assumption of independence when low correlation coefficients are extracted, as the fluctuations from true randomness can mask other systematic components.

Fig. 4.12 illustrates the reproducibility as a function of number of averaging measurements to extract FPUF. The percentage of partial program number variation of individually tested bits is plotted for both cases of dynamic and fixed block erase. On average, over 8% partial program number fluctuations existed for direct use of one-time extraction of FPUFs, and a significant number of bits have variations exceeding 20% of the original partial program numbers. By averaging 10 measurements to produce new FPUFs, bit-wise fluctuations can be reduced to below 4%, with few bits exceeding 10% difference.

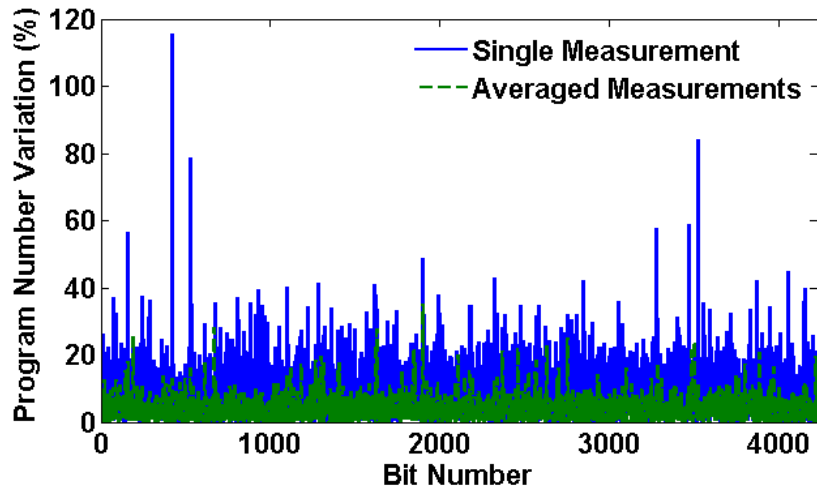


Figure 4.12 Percentage of partial program number variations of FPUF responses obtained between single measurements and between averaged measurements.

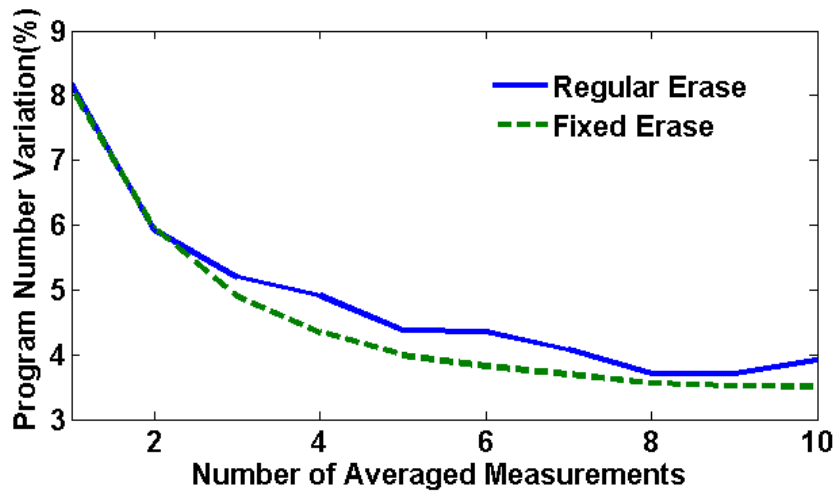


Figure 4.13 Average percentage bit-wise fluctuations between FPUFs with full erase and fixed erase against number of averaged measurements.

If averaging measurements are used to reduce FPUF variations, global erase can add undesirable bias between each group of measurements, since full erase times are dynamically adjusted and hard to control. Effects on FPUFs obtained by

conventional full erase and fixed-time erase are plotted in Fig. 4.13. Global erase biasing can reduce the consistency of the extracted FPUF due to its dynamic nature.

4.6 *Quantization Algorithms and Respective Maximum Entropies*

4.6.1 *Short-and-Long Bits*

Physical unclonable function (PUF) is a one-way function based on the intrinsic, uncontrollable but reproducible randomness of the implementing physical system. Various IC instances have been exploited to realize PUFs, which usually require custom circuits to obtain limited amount of output bits [5, 10, 19]. Flash memory based PUFs (FPUFs) [22, 23] have several practical advantages over other implementations, such as wide applicability and independency of power cycles [5]. Since Flash is one of the most aggressively scaled cells, FPUF is also superior in bit capacity compared to other PUFs [26]. The analog nature of the FPUF responses can also provide additional entropy compared to the digital outputs of prior PUF implementations. All of these unique characteristics of FPUFs are extremely attractive for embedded system security applications.

Unlike all prior PUF implementations, FPUF responses are unique in a sense that they are analog numbers with finite precision errors, which can be quantized in many different ways to generate crypto keys for various applications. Intuitively, partial program number (PPN) of an individual Flash cell can be used to generate a single PUF bit. A simple quantization algorithm is illustrated in Fig. 4.14.

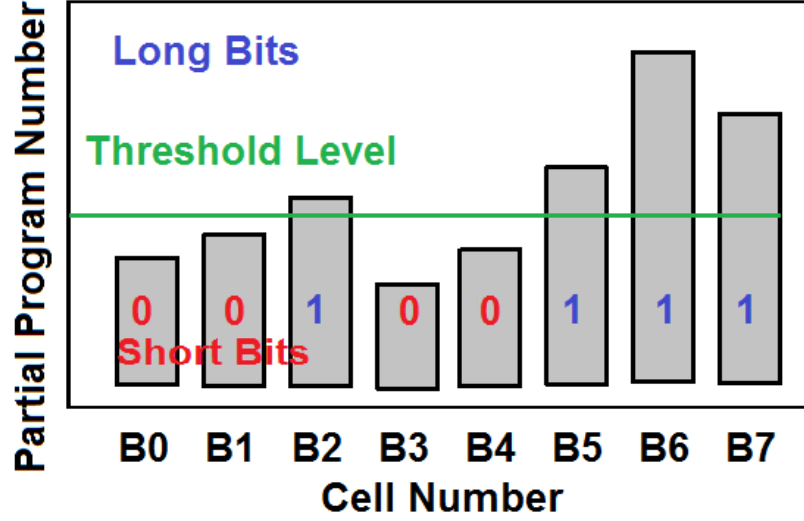


Figure 4.14 Simple short-long quantization algorithm for FPUF.

Flash cells requiring more partial program numbers than a certain threshold value can be thought as “long” bits with an assigned value “1”. Cells requiring less partial program numbers can be thought as “short” bits with digital value “0”. If N Flash cells are read sequentially to generate crypto keys, and the threshold level is set so that 0’s and 1’s are nearly equally distributed, the maximum entropy will simply be N bits. If permutation of Flash cells is allowed to generate keys [27], then the maximum entropy of the system will be $\log_2(N!)$. Since bit capacity per unit area of FPUF is orders of magnitude higher than other PUF implementations in the same technology [26], entropy of FPUF system can be superior with very large N .

4.6.2 Pair-wise Comparison

Due to random telegraph noise (RTN) induced variations during operations [26], cells with PPN close to threshold level may have large bit error rate and reduced overall FPUF consistency. A different way to generate independent bits is pair-wise comparison, which is similar to extracting arbiter [10] and ring oscillator PUFs [19]. If

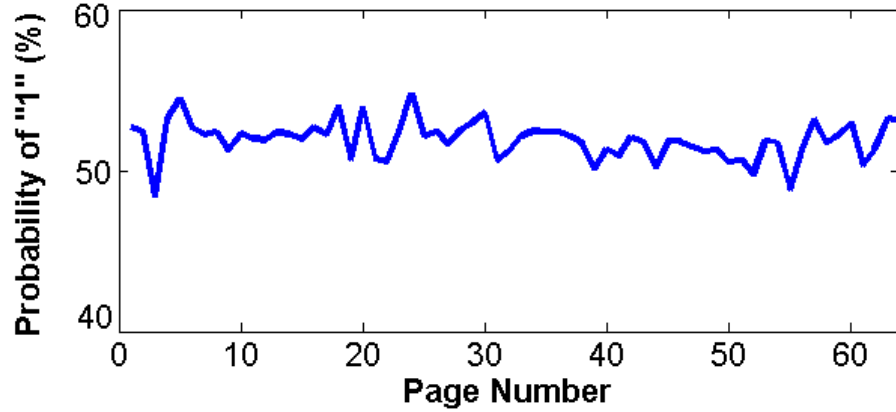


Figure 4.15 Probabilities of pair-wise comparison yield value “1” across all pages in a block.

Flash cell A requires higher PPN than adjacent Flash cell B, then the comparison will yield a “1”, otherwise the result will be “0”. Fig. 4.15 shows the experimental probability of the pair-wise comparison result equal to “1” over all pages in a block. The probability of randomly obtaining “0” or “1” is almost evenly distributed. This demonstrates each pair-wise comparison has entropy close to 1 bit. Similar distributions can be obtained from different cell pairings as well.

4.6.3 Cell-wise and Pair-wise Entropies

Since the FPUF response are extracted as analog numbers of PPN, the corresponding maximum cell-wise and pair-wise entropies can be much greater than 1 bit depending on different quantization methods. Fig. 4.16 (a) and (b) illustrate the probability mass functions (PMFs) of single and pair-wise Flash partial program number distributions respectively.

Corresponding entropies can be calculated from PMFs based on Shannon entropy [28]. On average, maximum single cell entropy is close to 9 bits and pair-wise

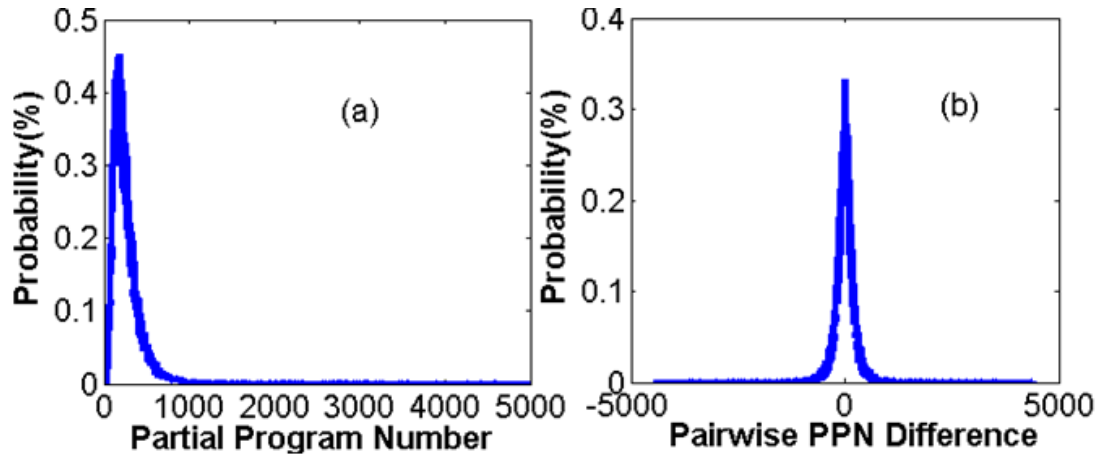


Figure 4.16 PMF for (a) single Flash cell (b) pair-wise partial program number distributions.

entropy is around 9.7 bits. Different quantization methods will reduce the cell-wise and pair-wise entropies accordingly, with lower bound close to 1 bit per cell or per pair.

4.7 *Future Possibility for Data Encryption*

Previously, FPUF is applied only to generating unique authentication signatures and challenge-response pairs [19, 22, 23]. Advantages in bit capacity, entropy and extraction method allow FPUFs to generate as many bits as plaintext with ease, unlike conventional binary PUFs. This denotes promises in a secured information coding scheme due to the analog characteristic of PPN, which permits various manipulations to the original FPUF responses, thus providing additional security to the encrypted information. A simple example is demonstrated in Table 4.4 just to show the potential of utilizing FPUF in a slightly different way, where neither perfect encryption algorithm nor superior security is assumed in this particular example.

Table 4.4 Simple example of utilizing quantized FPUF for data encryption.

Operations	Bit Numbers and Corresponding Values							
	B0	B1	B2	B3	B4	B5	B6	B7
Plaintext	1	1	1	1	0	0	0	0
Initial PPN	300	500	700	600	400	800	950	900
Original FPUF	L	L	L	L	L	L	L	L
New PPN	0	100	300	200	0	400	550	500
New FPUF	S	S	L	S	S	L	L	L
Crypto Key	0	0	1	0	0	1	1	1
Cipher Text	1	1	0	1	0	1	1	1
Reader Decipher	1	1	1	1	0	0	0	0
Attacker Decipher	0	0	1	0	1	0	0	0

FPUF outputs are first digitized through short-and-long quantization with threshold set at distribution median of PPN_{med} equals to 265. Long bits are chosen to be part of the pad, because additional manipulation can be performed to create the actual crypto keys in use. There are many ways to manipulate the initial FPUF, and a simple one is to apply additional partial program pulses to the pad bits before extracting the crypto key. In practice, which cells and how many additional PPN will be applied can be determined through password-generated hash functions, either based on symmetric keys or public keys. For the sake of demonstration, a universal 400 partial program pulses are applied to all pad bits, and new PPN can be obtained by the intended reader. The encrypted text is then created through XOR. With the correct crypto key, the reader can successfully retrieve the plaintext. This is by no means a secured or matured way to generate crypto keys using FPUFs, since simply applying 400 partial program pulses are easily detected, but just as a demonstration of the flexibility of FPUF. More secured and advanced algorithm is required for actual applications, which will belong to the future work.

Since the pad are generated based on FPUF, without obtaining the physical

Flash chip, attacker has no chance of retrieving the plaintext, just as OTP. If somehow the attacker obtained the Flash chip, the additional manipulation to the original FPUF can result in a 50% bit error rate, as demonstrated in the example.

4.8 Conclusion

This study illustrates the importance of characterizing the physical mechanisms behind FPUFs. FPUF physical origins are decomposed into two major categories: (1) manufacturing variations that include systematic and intrinsic random components; (2) field variations that come from RTN and aging and affect the reproducibility of FPUFs. Strong systematic variations can severely degrade the uniqueness and entropy of the FPUF bits, and should be properly removed in appropriate security applications. Bit-wise FPUF fluctuation is caused by inherent RTN during operations, and ways to improve FPUF designs in both uniqueness and reliability are discussed.. Erase fluctuation can add undesired global biases, but can be alleviated by fixing the erase time.

This study also demonstrates possible ways to utilize the analog responses extracted from FPUF for embedded security applications. Entropies associated with various quantization methods are discussed, and maximum bit-wise and pair-wise entropies are also given, which are significantly higher than the conventional digital outputs. Example of utilizing the generated FPUF bits with their analog nature is illustrated, which demonstrated the flexibility of FPUF for data encryption that is different with any other binary coding scheme. This unique analog nature of the response allows more extensive manipulations to the original FPUF, potentially creating additional information security.

REFERENCES

- [1] A. Asenov, “Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 micron MOSFETs: A 3-D “atomistic” simulation study”, *IEEE Trans. Electron Devices*, vol. 45, pp.2505 -2513 1998.
- [2] A. Blum, A. Kalai and H. Wasserman, “Noise-tolerant learning, the parity problem, and the statistical query model”, *J. ACM*, vol. 50, no. 4, pp. 506 – 519, 2003.
- [3] E. Chang, B. Stine, T. Maung, R. Divecha, D. Boning, J.Chung, K. Chang, G. Ray, D. Bradbury, O. S. Nakagawa, S. Oh, and D. Bartelink, “Using a statistical metrology framework to identify systematic and random sources of die- and wafer-level ILD thickness variation in CMP processes”, *IEDM Tech. Dig.*, Dec. 1995.
- [4] B. Gassend, D. E. Clarke, M. van Dijk, and S. Devadas, “Silicon Physical Unknown Functions”, *ACM Conference on Computer and Communications Security*, 2002.
- [5] J.Guajardo, S. S. Kumar, G.-J. Schrijen, and P. Tuyls, “FPGA intrinsic PUFs and their use for IP protection”, *Workshop on Cryptographic Hardware and Embedded Systems 2007*, Sep. 2007.
- [6] D. E. Holcomb, W. P. Burleson, and K. Fue, “Initial SRAM State as a Fingerprint and Source of True Random Numbers for RFID Tags”, *Proc. of the Conference on RFID Security*, Jul. 2007.
- [7] S. Kamohara, D. Park, and C. Hu, “Deep-trap SILC model for nominal and weak oxides”, *Proc. IRPS*, 1998.

- [8] P. Koeberl, R. Maes, V. Rozic, V. Van der Leest, E. Van der Sluis, and I. Verbauwhede, “Experimental evaluation of Physically Unclonable Functions in 65 nm CMOS”, *IEEE European Solid-State Circuit conference*, Sep. 2012.
- [9] S. S. Kumar, J. Guajardo, R. Maes, G. J. Schrijen, and P. Tuyls, “The Butterfly PUF: Protecting IP on every FPGA”, *IEEE International Workshop on Hardware-Oriented Security and Trust 2008*, 2008.
- [10] Jae W. Lee, D. Lim, B. Gassend, G. E.Suh, M. van Dijk, and S. Devadas, “A technique to build a secret key in integrated circuits for identification and authentication application”, *Symposium on VLSI Circuits*, 2004.
- [11] R. Maes, P.Tuyls, and I. Verbauwhede, “Intrinsic PUFs from flip-flops on reconfigurable devices”, *Workshop on Information and System Security*, 2008.
- [12] G. Marsaglia, *The Marsaglia random number CDROM including the Diehard battery of tests of randomness*, Florida State University, 1995.
- [13] Open NAND Flash Interface Specification. Hynix Semiconductor, Micron Technology.
- [14] R. Pappu, B. Recht, J. Taylor, and N. Gershenfeld, “Physical one-way functions”, *Science*, vol. 297, no. 5589, 2002.
- [15] P. Prabhu, A. Akel, L. Grupp, W. Yu, G. E. Suh, E. Kan and S. Swanson, “Extracting device fingerprints from Flash memory by exploiting physical variations”, *Trust and Trustworthy Computing, Lecture Notes in Computer Science*, Springer Berlin / Heidelberg, 2011.
- [16] Samsung Electronics. 512M x 8 Bit/1G x 8 Bit NAND Flash Memory. K9F4G08U0A datasheet. Nov. 2005 [Revised Jan. 2006].

- [17] P. Simons, E. van der Sluis, and V. van der Leest, "Buskeeper PUFs, a promising alternative to D Flip-Flop PUFs", *IEEE Int.Symposium on Hardware-Oriented Security and Trust*, 2012.
- [18] Y. Su, J. Holleman, and B. Otis, "A 1.6pJ/bit 96% stable chip-ID generating circuit using process variations", *IEEE Int. Solid-State Circuits Conference*, Feb. 2007.
- [19] G. E. Suh and S. Devadas, "Physical unclonable functions for device authentication and secret key generation", *Design Automation Conference*, 2007.
- [20] M. J. Uren, D. J. Day and M. J. Kirton, "1/f and random telegraph noise in silicon metal-oxide-semiconductor field-effect transistors", *Appl. Phys. Lett.*, vol. 47, no. 11, pp.1195 -1197, 1985.
- [21] V. Vivekraj and L. Nazhandali, "Circuit-level techniques for reliable physically uncloneable functions", *IEEE International Workshop on Hardware-Oriented Security and Trust*, Jul. 2009.
- [22] Y. Wang, W. Yu, G. E. Suh, and E. Kan, "Flash memory for ubiquitous hardware security functions: true random number generation and device fingerprints", *Proc. of the IEEE Symposium on Security and Privacy*, 2012.
- [23] Y. Wang, W. Yu, S.Q. Xu, E. Kan , G. E. Suh, "Hiding information in Flash memory", *Proc. of the IEEE Symposium on Security and Privacy*, 2013.
- [24] H. S. Wong and Y. Taur, "Three dimensional 'atomistic' simulation of discrete random dopant distribution effects in sub-0.1 μ m MOSFET's", *IEDM Tech. Dig.*, 1993.

- [25] Y. Ye, F. Liu , S. Nassif , Y.Cao, “Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness”, *Proc. of the 45th annual Design Automation Conference*, June, 2008
- [26] S.Q. Xu, W. Yu, G. E. Suh, and E. Kan, “Understanding sources of variations in Flash Physical Unclonable Function”, *Submitted to DAC* 2014.
- [27] E. En Gad, E. Yaakobi, A. Jiang, J. Bruck, “Rank-modulation rewriting codes for flash memories”, *Proc. of the ISIT*, 2013.
- [28] P. Tuyls, B. Skoric, S. Stallinga, A.H.M. Akkermans, and W. Ophey, “Information-theoretic security analysis of Physical Uncloneable Functions”, *Financial Cryptography and Data Security*, Springer Berlin Heidelberg, 2005.

CHAPTER 5

PROBING THE ORBITAL LEVELS OF ENGINEERED FULLERENIC MOLECULES FROM A NONVOLATILE MEMORY CELL

5.1 *Abstract*

The Coulomb blockade behavior was observed for both C60-PCBM and C70-PCBM at room temperature utilizing a nonvolatile memory cell fabricated through a liquid-transfer process. Room-temperature and low-temperature (10K) electrical characterizations verified the blockade effect was originated from both molecular energy levels and single electron charging energy. Molecular orbital energy was extracted and shown good agreement with the literature [1]. The successful integration and operation of this hybrid structure signified a strong potential for molecule-based electronic device design. The mono-disperse nature of the molecules is the natural way to eliminate inherent process variations.

5.2 *Introduction*

Engineered fullerene molecules (EFM) are chemical derivatives of neat fullerene molecules with multiple functionalities. The mono-dispersed nanoscale size of EFM brings forth improved scalability and reduced device variations. The redox capability [2] and electrical conductivity [3] of EFM are notably different from pristine fullerenes and offer more flexibility in tailoring the fabrication process and device characteristics. Most importantly, chemical functionalization in EFM alters the electronic structure of the molecule, creating programmable HOMO-LUMO levels [4] which are crucial for designing resonant tunneling barrier for Flash memory to overcome the scaling bottleneck [5]. This chemical derivation also grants EFM large

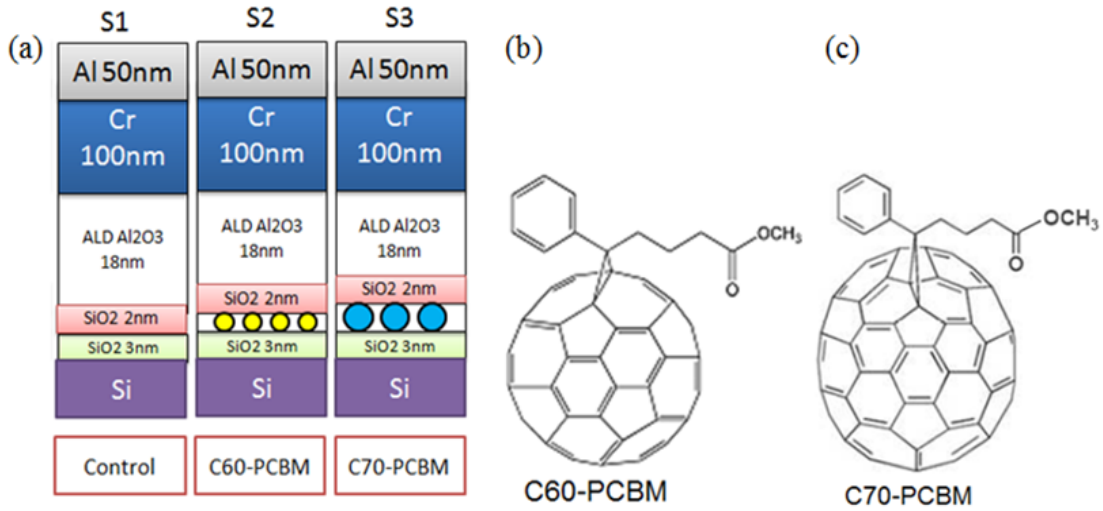


Figure 5.1 (a) Design splits of the EFM gate stack: S1 is the control sample, S2 contains C60-PCBM floating gate, and S3 contains C70-PCBM floating gate. (b) Chemical structure of the C60-PCBM molecule. (c) Chemical structure of the C70-PCBM molecule.

solubility at room temperature that enables wafer-level fluid-transfer process, which may ease both the process control and manufacture cost in the case of commercialization.

5.3 Experiments

A Metal-Oxide-Semiconductor (MOS) capacitor structure in the gate stack of conventional flash memory cell [6] was fabricated and electrically characterized. The experimental splits are listed in Fig. 5.1. The device was fabricated on top of a 4" p-type silicon substrate, with a 3nm thermally grown tunnel oxide. Floating gates for charge storage including design splits of two different EFM species were formed on top of the tunnel oxide.

All EFMs used in the experiment were chemically synthesized at Nano-C Inc. and received as powder form with over 99% purity. Sample S1 is the control device

without any embedded EFM. Samples S2 and S3 integrate fluid-transferred C₆₀-PCBM and C₇₀-PCBM through the room-temperature spin-coating method using toluene as the solvent. The pristine C₆₀ and C₇₀ are not included in the control samples due to the low solubility.

The initial spin-coating recipe was developed based on the model of Newtonian liquid on a rotating disk [7], and later improved through experimental trials. The targeted concentration was calculated to be 0.5 mg/ml based on an estimated molecule number density of 10^{12} cm^{-2} as the floating gate layer [5]. The EFM solution was prepared in the glove box with Ar ambient, and the spin-coating process was performed subsequently at the speed of 2000 rpm. The toluene residuals were then cleaned through evaporation in a nitrogen purge ambient at 250°C for 15 hours. An evaporated SiO₂ of 2nm is deposited to protect the EFM and to enable the atomic layer deposition (ALD) of the Al₂O₃ control dielectric with 18nm thickness, confirmed by Woollam spectroscopic ellipsometer. E-gun evaporated 100nm Cr and 50nm Al were used as metal gate and patterned through wet chemical etch. Reactive ion etching is then used to pattern the remaining gate stack to avoid formation of a large-area floating gate (FG) across the entire wafer. Finally, a 400°C annealing was performed in the forming gas for 30min to passivate the interface and enhance metal conductivity.

5.4 Discussion

High-frequency (1MHz) capacitance-voltage (CV) measurements were performed on all samples by a Keithley 590 CV analyzer at both room temperature and low temperature (10K). The extracted flat-band voltage shift (ΔV_{FB}) versus programming voltage (V_{Program}) is illustrated in Fig.5. 2 (a) and (b) for C₆₀-PCBM and C₇₀-PCBM respectively.

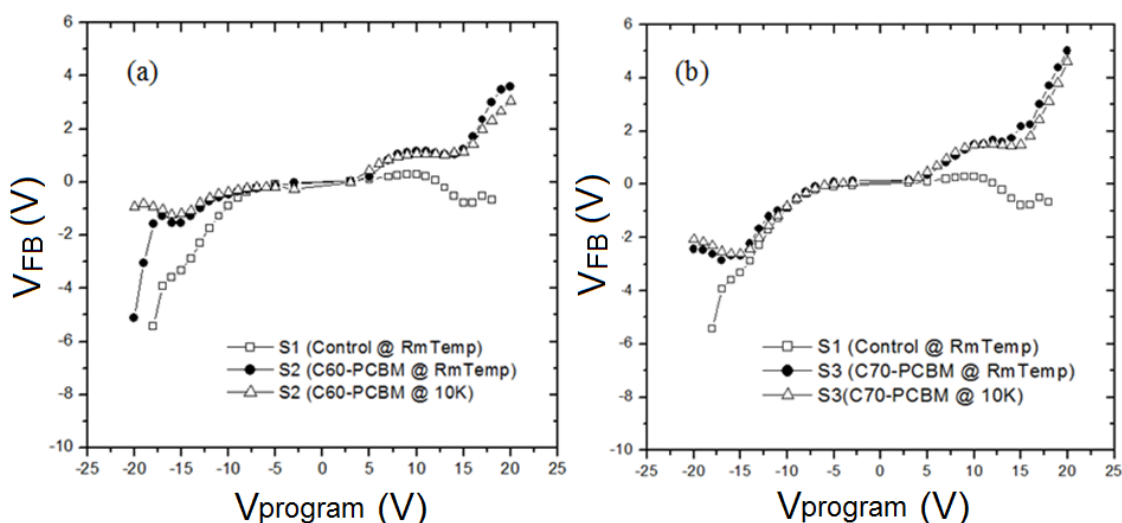


Figure 5.2 Room-temperature and low-temperature (10K) high frequency CV measurements on toluene-spin-coated (a) C₆₀-PCBM and (b) C₇₀-PCBM in comparison with the control sample S1.

Discrete charge injection manifests as ΔV_{FB} plateaus and can be discerned clearly for both C₆₀-PCBM and C₇₀-PCBM; minimal difference is seen between the two temperatures, proving that the Coulomb staircase behaviour observed at room temperature did not arise from Frenkel-Poole (F-P) conduction in the control dielectric [5], but rather due to the interaction of the LUMO levels of the EFM in the program operation. This is also confirmed by the low density of the interface and trap states shown in the control sample S1, where only negligible ΔV_{FB} is present up to 12V. The flatness of the plateaus in Fig. 5.2, representing minimal energy dispersion, is a strong indication that the solution-phase protocol did keep the integrity of the EFM molecules. Otherwise disintegrated molecules would produce a continuous increase of V_{FB} , dictating a distribution of energy levels. Similar results were obtained in the case of C₇₀-PCBM as seen from Fig. 5.2 (b).

The Coulomb staircase or blockade behavior originates from both the

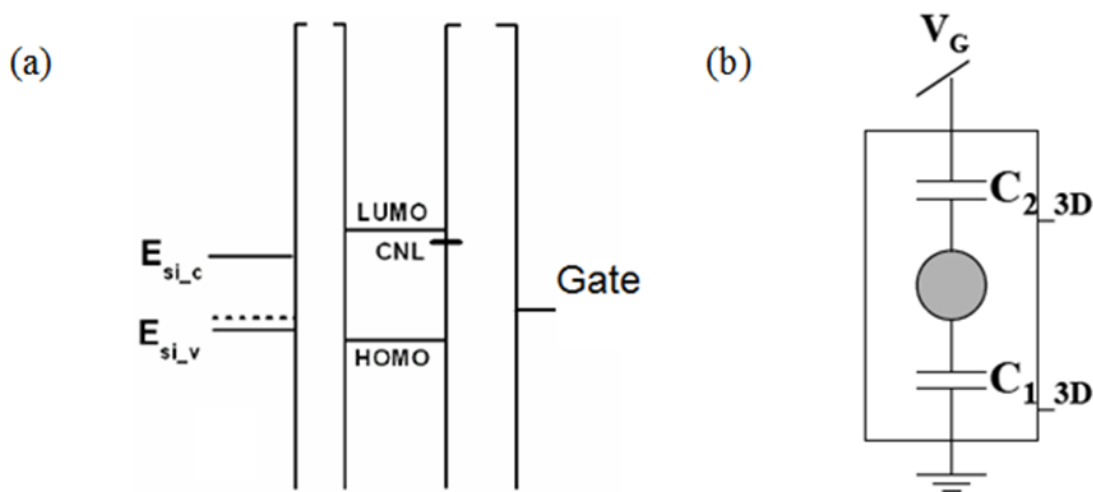


Figure 5.3 (a) Band diagram and (b) schematic of the MOS capacitor gate stack [8, 9].

electronic states of the EFM and the self capacitance (C_{FG}) of the molecules due to their small sizes. The band diagram and the unit cell schematic are portrayed in Fig. 5.3.

The capacitive coupling among the gate, the molecule and the substrate are modeled by a series of capacitors similar to the conventional continuous floating-gate devices. When the gate voltage (V_G) is swept, the potential on the embedded EFMs (ΔE_{EFM}) also changes accordingly. Adjusting the EFM potential is equivalent to shifting the energy levels of the EFMs in the band diagram, and can be thought as sweeping the EFM energy levels relative to the Si conduction band edge E_{Si_C} . The coupling capacitance $C1_3D$ and $C2_3D$ are established using the three-dimensional (3D) electrostatic model [8, 9].

In order to initiate electron injection into the lowest available EFM energy level--LUMO (lowest unoccupied molecular orbital), ΔE_{EFM} should be large enough to overcome the offset between E_{Si_C} and the LUMO level. Secondly, the LUMO level need to be moved further to give the extra electron charging energy (E_{CH}) difference

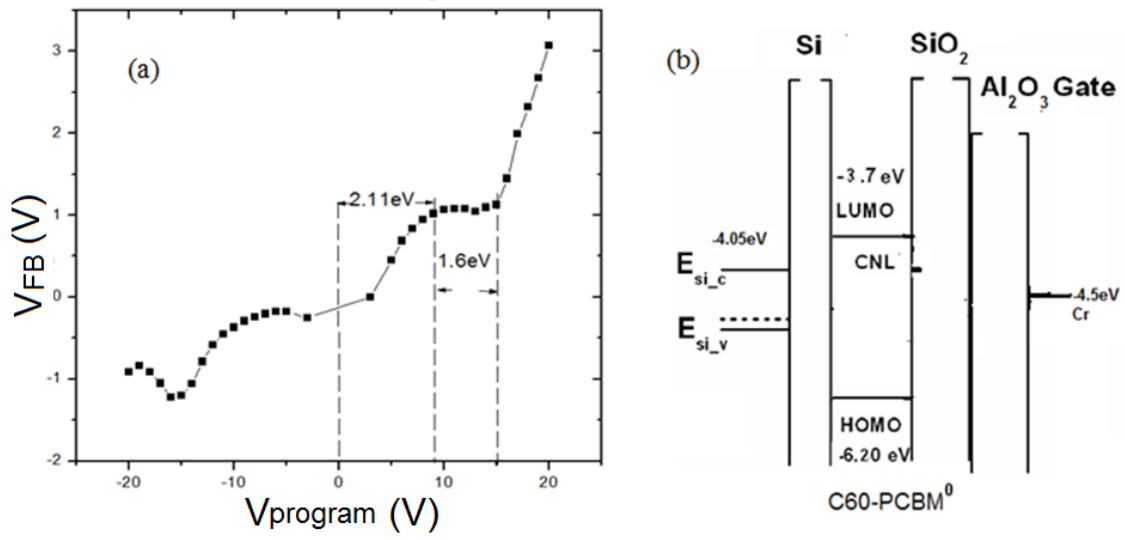


Figure 5.4 (a) ΔE_{EFM} extraction from ΔV_{FB} for C₆₀-PCBM. (b) The band diagram from calculated C₆₀-PCBM molecular orbital.

required for the electron to inject into the LUMO level. The electrostatic energy E_{CH} is approximated through the 3D model. These relationships can be summarized in the following equations.

$$\Delta E_{EFM} = \text{Band offset between } E_{Si_C} \text{ and LUMO} + E_{CH} \quad (5.1)$$

$$E_{CH} = e^2 / C_{FG} \quad (5.2)$$

$$C_{FG} = C1_3D + C2_3D \quad (5.3)$$

From the experimental data in Fig. 5.4, the starting point of the Coulomb plateau represents the initiation of the electron injection. By extracting the ΔE_{EFM} and calculating the E_{CH} , an estimation of the LUMO level of the EFM's can be obtained.

For illustration, ΔE_{EFM} required for entering the first plateau of C₆₀-PCBM was extracted from the low temperature measurements to be 2.11eV as illustrated in Fig. 5.4 (a). For the 3D electrostatic model, molecules were modeled as rigid metal spheres in a 2D lattice. The unit cell area, which directly related to the number density (N), was extracted through ΔV_{FB} , as depicted in Equations (5.4) and (5.5).

$$\Delta V_{\text{FB}} = k \cdot \frac{C_{1-3D}}{C_{1-1D}} \frac{e \cdot Q}{C_{2-3D}} = k \cdot \left(\frac{C_{1-3D}}{C_{1-1D}} / \frac{C_{2-3D}}{C_{2-1D}} \right) \frac{e \cdot Q}{C_{2-1D}} = R_{3D} \cdot \frac{e \cdot Q}{C_{2-1D}} \quad (5.4)$$

$$C_{2-1D} = \frac{\varepsilon_{\text{conl}}}{T_{\text{conl}}} \cdot A = \frac{\varepsilon_{\text{conl}}}{T_{\text{conl}}} \cdot \frac{1}{N} \quad (5.5)$$

k is a correction constant translating the capacitance ratio to ΔV_{FB} , which is usually very close to unity; R_{3D} is the 3D channel-control factor, A is the unit cell area, T_{conl} is the thickness of the control dielectric and $\varepsilon_{\text{conl}}$ is the permittivity of the control dielectric[8, 9].

For C₆₀-PCBM, the molecular diameter was chosen within a range of 0.7~1.2nm to encompass both the physical radius of C₆₀ and the reported van der Waal radius of C₆₀-PCBM [10]. The corresponding C_{FG} in the gate stack was roughly 0.56~0.6(e/V), which gave an E_{CH} around 1.6~1.7eV. This number is the single electron charging energy and corresponds well with the duration of the first plateau (Fig. 5.4 (a)).

Due to the extremely low density of the interface state, Fermi-level pinning due the charge neutrality level (CNL) [11] may not be significant, and the molecules were assumed to be initially neutral. Therefore, according to Equation (1), the LUMO level of the C₆₀-PCBM was determined to be -3.7eV (Fig. 5.4 b), matching the published data very well [1]. For C₇₀-PCBM, similar calculations were carried out

with the effective diameter of the molecule between 0.8~1.4nm, which corresponds to an E_{CH} around 2.3~2.4eV. The ΔE_{EFM} was extracted to be 2.27eV, which gave the LUMO level of C70-PCBM around -4.0eV. This lower LUMO level observed in C₇₀-PCBM is consistent with previous studies on the electronic structures of their pristine counter parts, where C₇₀ has a lower LUMO level compared to C₆₀ [12].

5.5 *Conclusion*

We A nonvolatile memory cell was employed in this study to evaluate the molecular energy levels of the selected EFM species. Room-temperature Coulomb staircase was successfully demonstrated for both C₆₀-PCBM and C₇₀-PCBM with very low interface trap density in the fluid-transfer process. The extracted LUMO levels corresponded well with literature [1]. This investigation paves way for a nondestructive method to characterize the electronic structure of molecules as well as fluid-transfer process integration for hybrid molecular electronic devices.

REFERENCES

- [1] C-W. Chu, V. Shrotriya, G. Li and Y. Yang, "Tuning acceptor energy level for efficient charge collection in copper-phthalocyanine-based organic solar cells", *Appl. Phys. Lett.* 88, 153504, 2006.
- [2] C. Lee, U. Ganguly, and E. C. Kan, "Characterization of number fluctuations in gate-last metal nanocrystal nonvolatile memory array beyond 90nm CMOS technology", *MRS Proceedings*. Vol. 830. No. 1, Boston, MA, Nov. 29 – Dec. 3, 2004.
- [3] I. I. Mazin, S.N. Rashkeev, V.P. Antropov, O. Jepsen, A. I. Liechtenstein and O.K. Andersen, "Quantitative theory of superconductivity in doped C60", *Phys. Rev. B* 45, pp.5114-5117, 1992.
- [4] B. M. Illescas, N. Martin and C. Seoane, "Reaction of C60 with Sultines: Synthesis, Electrochemistry, and Theoretical Calculations of Organofullerene Acceptors", *J. Org. Chem.*, 62, no.22, pp.7585–7591, 1997.
- [5] T.-H. Hou, H. Raza, K. Afshari, D. J. Ruebusch and E. C. Kan, "Nonvolatile memory with molecule-engineered tunneling barriers", *Appl. Phys. Lett.* Vol. 92, 153109, 2008.
- [6] C. Lee, J. Meter, V. Narayanan and E. C. Kan, "Process characterization of metal nanocrystal self-assembly on ultra-thin oxide for nonvolatile memory applications", *J. Semiconductor Materials*, vol. 34, no. 1, pp.1-11, 2005.
- [7] D. B. Hall, P. Underhill and J. M. Torkelson, "Spin coating of thin and ultrathin polymer films", *Polymer Engineering and Science*, vol.38, no.12, pp.2039-2045, 1998.
- [8] T.-H Hou, C. Lee, V. Narayanan, U. Gangly and E. C. Kan, "Design optimization of metal nanocrystal memory—Part I: Nanocrystal array engineering", *IEEE Trans. Electron Devices*, vol.53, pp.3095-3102, 2006.

- [9] T.-H. Hou, C. Lee, V. Narayanan, U. Ganguly and E. C. Kan, "Design optimization of metal nanocrystal memory—Part II: Gate-stack engineering", *IEEE Trans. Electron Devices*, vol.53, pp. 3103-3109, 2006.
- [10] H. Tanaka and K. Takeuchi, "Diameter determination of C60 and C70 monomers in the gas phase using a differential mobility analyzer", *Appl. Phys. A* 80, pp.759-761, 2005.
- [11] J. Robertson, "Band offsets of wide-band-gap oxides and implications for future electronic devices", *J. Vac. Sci. Technol. B* 18, pp.1785–1791, 2000.
- [12] S. J. Woo, E. Kim, and Y. H. Lee, "Geometric, electronic, and vibrational structures of C50, C60, C70, and C80", *Phys. Rev. B* 47, 6721, 1993.

CHAPTER 6

CONCLUSION

6.1 *Summary of Major Contributions*

The major contributions of the work described in this dissertation are summarized as follows:

1. The severity of process variations on scaled SRAM is demonstrated through careful examinations of the RDF effects on a prototype 22nm SRAM. Monte Carlo technique is incorporated to establish a more realistic doping profile, and 3D atomistic simulations are performed to generate individual transistor response as well as SNM and SINM. The study also confirms that replicating the RDF induced electrical characteristic on a large scale requires extensive computational power.
2. A hybrid SRAM-DRAM cell with cross capacitor is proposed to provide both multi-bit storage capability, mismatch tolerance, and disturb stabilization to mitigate the severe SRAM scaling challenges concluded from previous studies.
3. CMOS variability sources are decomposed and characterized to provide theoretical foundations for better implementation of FPUF, which is a unique way to utilize inherent process variation in Flash memory for security applications.
4. Engineered fullerenic molecules are explored to understand their energy states when embedded in NVM gate stack, paving way towards hybrid molecular integration with tunable tunneling barrier. The inherently mono-dispersed molecule size brings forth reduced device variations.

6.2 *Suggestions for Future Work*

6.2.1 *Flash Rank Modulation*

Besides the additional work mentioned at the end of Chapter 4 on finding a suitable algorithm to utilize the full analog nature of the FPUF for data encryption, potentials on exploiting partial program operation, which is a unique way to show device-level characteristics without any additional hardware, may pave way towards implementing a very interesting idea, called Flash rank modulation [1].

Conventional Flash memory data are represented by the absolute threshold voltage values. As illustrated in Fig.6.1, single-level cell (SLC) has larger memory window between the two distinctive states, allowing faster access time and availability

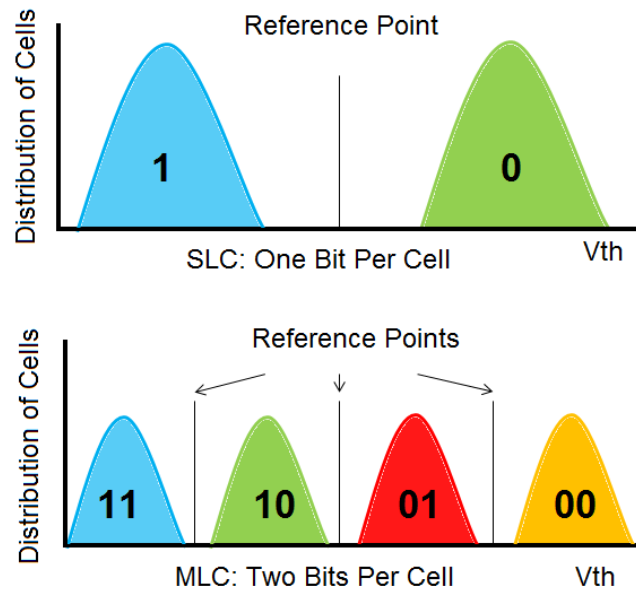


Figure 6.1 Conventional Flash memory programmed states for SLC and MLC devices.

Table 6.1 Relationship between group size N and resulted equivalent bit capacity [1].

Group Size N	2^N	N!	$\text{Log}_2 N!$
2	4	2	1
3	8	6	2.6
4	16	24	4.6
5	32	120	6.9
6	64	720	9.5
7	128	5,040	12.3
8	256	40,320	15.3
9	512	362,880	18.5

for embedded applications, but can only store one bit per cell. Multi-level cell (MLC) on the other hand, can store two or more bits per cell, but suffers from reduced access speed due to requirement on controlling the accuracy and reliability caused by the reduced memory windows [2]. Many performance optimizing codes have been proposed based on the conventional data representation in hopes of achieving higher bit capacity without the penalties on accuracy and reliability [3-6].

Recently, a new coding scheme that changes how data is fundamentally represented in Flash memory has been proposed, which is called Flash rank modulation [1,7]. Instead of using absolute V_{th} values to represent data, rank modulation uses the relative order of cell levels from a group to convey information. This method does not only help with overshoot problems during MLC program operation, but also relax the necessity of having global erase when altering the stored data [8]. For example, if a group of 3 cells are used to store a single information packet, six possible permutations can be represented: (1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2) and (3, 2, 1). Therefore, a 3-cell group can store equivalent of $\log_2 6 \approx 2.6$ binary bits. Because the data is represented in a relative format, the absolute V_{th}

values for all the cells are not important as long as their relative ranks are maintained, which makes this method much more tolerant to process and field variations. In addition, if different data need to be stored in these 3 cells, only additional program operation is required to raise the relative level of the desired cells, instead of doing a global erase that is costly in both energy and time [1]. The third advantage of rank modulation is its potential in increasing data capacity per cell. Table 6.1 lists the relationship between group size N and equivalent bit capacity $\log_2 N!$ for $N < 10$; results with $N \geq 10$ can be calculated through Sterling approximation.

6.2.2 Trial Implementation with Partial Program Operation

In literature, rank modulation scheme is still in a very primitive stage [1, 8, 9], and suggestions on its implementation usually require elaborated custom circuitry [10]. However, the partial program operation mentioned in Chapter 4 may shine some light on this topic. Preliminary analysis and experimental results are summarized below; the experiment was proposed to implement rank modulation in SLC to improve data

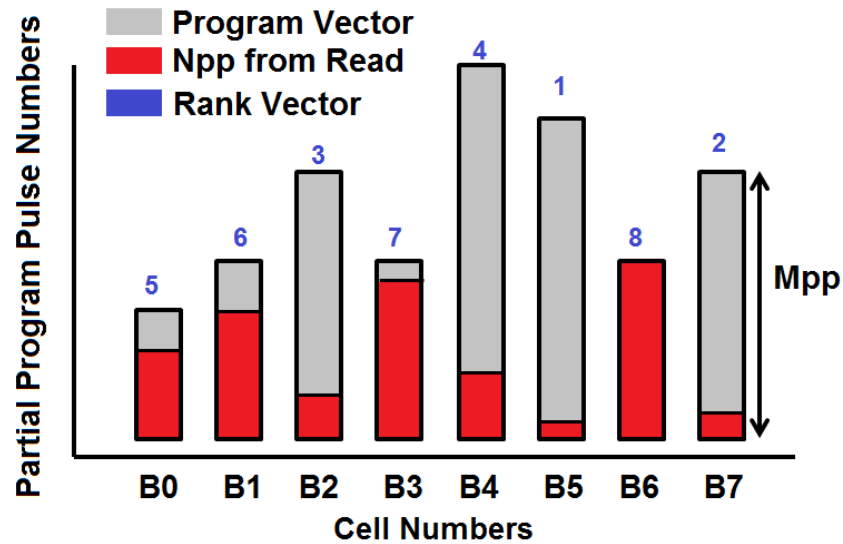


Figure 6.2 SLC rank modulation algorithm utilizing partial program operations.

Table 6.2 Sample program and read operation results for rank modulation.

Op. No.	Operations	B0	B1	B2	B3	B4	B5	B6	B7
1	Dynamic Mpp Extraction	388	493	723	497	1181	849	335	756
2	Desired Rank Vector	5	6	3	7	4	1	8	2
3	Dynamically Calculated PV	178	241	597	203	1013	807	0	672
4	Expected Read Npp	210	252	126	294	168	42	335	84
5	Actual Read Npp Obtained	200	264	144	294	46	1	357	141
6	Actual Rank Obtained	5	6	4	7	2	1	8	3

capacity in embedded applications without any custom circuitry. The same commercial off-the-shelf testing board mentioned in Chapter 4 is utilized in this setup. Group of 8 cells (B0, B1, B2, B3, B4, B5, B6, B7) are used to represent a single rank formation, and theoretically can represent 15.3 binary bits, which is close to MLC bit capacity. The generic program and read operations are illustrated in Fig. 6.2, and preliminary algorithms are explained below.

Due to process variations mentioned in Chapter 1, each Flash cell requires different number of partial program pulses to be switched from erased state “1” to programed state “0”. This initial set of partial program numbers for all 8 cells is called Mpp = (Mpp0, Mpp1, Mpp2, Mpp3, Mpp4, Mpp5, Mpp6, Mpp7). Program operation is achieved by initially partial program 8 cells with a set of numbers called program vector (PV) = (P0, P1, P2, P3, P4, P5, P6, P7). PV is pre-determined by the rank vector (RV) = (5, 6, 3, 7, 4, 1, 8, 2), which represents the desired relative ranks between cells. The stored rank is then read out as the bit switch sequence, represented by a set of additional partial program pulses called Npp = (Npp0, Npp1, Npp2, Npp3, Npp4, Npp5, Npp6, Npp7). Smaller Npp meaning the cell require fewer partial program pulses, therefore switches quicker than cells with larger Npp. Some preliminary experiment results are shown below in Table 6.2.

Unfortunately, the resulted bit error rate (BER) was not acceptable for every day computation. 200 additional experiments were performed and the readouts are plotted in Fig. 6.3. It is observed that, although on average the read results represent the correct rank, due to the high level of dynamic fluctuations mentioned in Chapter1, such as RTN, the BER for a single read is too much for a reliable read. However, if new algorithms or scheme can be developed in the future, rank modulation is still a very promising research direction.

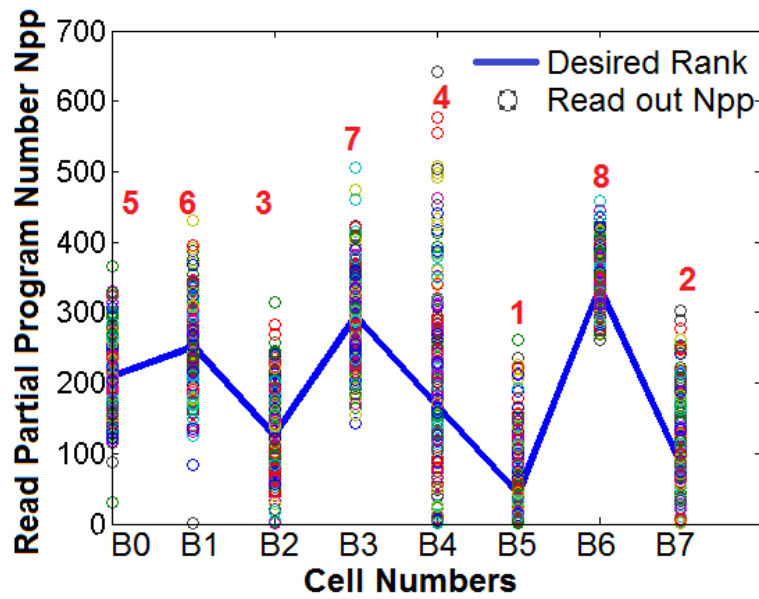


Figure 6.3 Read results for rank vector = (5, 6, 3, 7, 4, 1, 8, 2) with 200 samples.

REFERENCES

- [1] A. Jiang, R. Mateescu, M. Schwartz and J. Bruck, "Rank modulation for flash memories", *Proc. IEEE Int. Symp. Information Theory (ISIT)*, pp.1731 - 1735, 2008.
- [2] B. Eitan and A. Roy, "Binary and multilevel flash cells", *Flash Memories*, P. Cappelletti, C. Golla, P. Olivo, and E. Zanoni, Eds. Kluwer, pp. 91–152, 1999.
- [3] A. Jiang, "On the generalization of error-correcting WOM codes", *Proc. IEEE International Symposium on Information Theory (ISIT'07)*, pp. 1391-1395, 2007.
- [4] A. Jiang, V. Bohossian and J. Bruck, "Floating codes for joint information storage in write asymmetric memories", *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 1166-1170, 2007.
- [5] A. Jiang and J. Bruck, "Joint coding for flash memory storage", *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 1741-1745, 2008.
- [6] A. Jiang, M. Langberg, M. Schwartz and J. Bruck, "Universal rewriting in constrained memories", *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2009.
- [7] A. Jiang, M. Schwartz and J. Bruck, "Error-correcting codes for rank modulation", *Proc. IEEE International Symposium on Information Theory (ISIT)*, pp. 1736-1740, 2008.
- [8] E. En Gad, E. Yaakobi, A. Jiang and J. Bruck, "Rank-modulation rewriting codes for Flash memories", *Proc. IEEE International Symposium on Information Theory (ISIT)*, 2013.
- [9] A. Jiang and Y. Wang, "Rank modulation with multiplicity", *Proc. IEEE GLOBECOM Workshop*, pp.1866 -1870, 2010.

- [10] M. Kim, J. K. Park, and C. M. Twigg, “Rank modulation hardware for Flash memories”, *International Midwest Symposium on Circuits and Systems*, pp. 294 – 297, August 2012.